Protein inference in shotgun proteomics

5

5.1 Introduction to proteomics

Proteins are the key functional entities in the cell. Proteomics is the global analysis of proteins, which is critical to understanding how cells function. However, it is more challenging to gather information at the proteome level than at the genome and transcriptome levels [1]. This is because the proteome is complemented by alternative splicing and diverse posttranslational modifications. Meanwhile, proteins are interconnected with each other in the form of complexes and signaling networks that are highly divergent in time and space.

Mass spectrometry (MS) plays important roles in proteome analysis. With rapid developments in instrumentation, sample separation, and computational analysis, MS-based proteomics has been successfully used to characterize almost complete proteomes in a high-throughput manner [2]. Such maturation of MS-based proteomics will deliver answers to some important biological questions.

MS-based shotgun proteomics is a strategy that offers fast, high-throughput characterization of complex protein mixtures. In the experiments, the extracted proteins from the sample are first digested into peptides with protease such as trypsin. Enzymatic digestion of a full proteome can generate hundreds of thousands of peptides, making it unfeasible to perform the MS analysis directly. Hence, the liquid chromatography (LC) separation usually is first used to reduce the sample complexity before the MS analysis. Ideally, all peptides eluted from the LC should be captured by the mass spectrometer. However, this is not true since peptides compete for efficient ionization. Therefore, abundant peptides are more likely to be analyzed by the mass spectrometer than those less abundant peptides. As a result, not all peptides are captured by the mass spectrometer.

After ionization, peptide precursor ions are introduced into the mass spectrometer, which records both their mass-to-charge (m/z) ratio and intensity. The single-stage mass spectrum that is composed of peaks corresponding to peptide precursor ions is insufficient for ambiguous protein identification. Therefore, some single precursors are selected for further fragmentation to generate tandem mass spectra (MS/MS). In a tandem mass spectrum, there are generally two types of peaks: peaks generated from amino-terminal fragment ions ("*b*" ions) and peaks generated from carboxy-terminal fragment ions ("*y*" ions). The combination of precursor m/z and its tandem mass spectrum is used to determine peptide sequences, and then proteins are inferred from the identified peptides.

Finally, peptides and proteins are quantified (either relatively or absolutely) to generate protein abundance. These protein abundances are then interpreted and further used for biomarker discovery or protein–protein interaction network construction.

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00005-3

5.2 Protein identification in proteomics

In shotgun proteomics, the computational procedure for protein identification has two main steps: peptide identification and protein inference. In peptide identification, we search the experimental tandem mass spectra against a protein sequence database to obtain a set of peptide-spectrum matches, or use the *de novo* sequencing to determine the peptide sequences without using the protein database. In protein inference, those identified peptides are assembled into a set of confident proteins. Figure 5.1 gives an illustration of the protein identification process. In this chapter, we focus on the protein inference problem.

5.3 Protein inference: Problem formulation

Computationally, the input for the protein inference problem is a bipartite graph: one set of nodes is composed of identified peptides and another set of nodes is composed of candidate proteins that have at least one constituent peptide [3]. The inference problem considered here is to find a subset of proteins that are actually present in the sample. To date, many computational approaches for protein inference have been proposed. The details of these existing methods and the challenges of protein inference problem are summarized and discussed in Ref. [4].

The protein inference problem has been investigated from different perspectives. For instance, Ref. [3] has recently formulated it as a linear programming problem. This chapter focuses on how to use data mining techniques to solve this problem and shows that it can be tackled with several different data mining formulations.

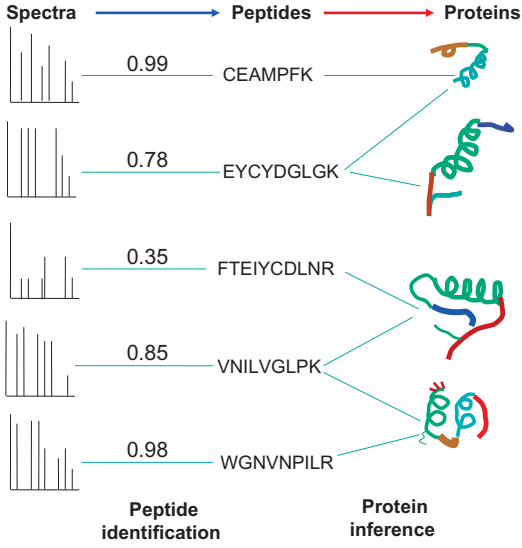


Figure 5.1 The protein identification process. In shotgun proteomics, the protein identification procedure has two main steps: peptide identification and protein inference.

5.4 Data collection

To infer proteins from identified peptides, the peptide identification results from the raw MS data and the corresponding protein sequence database should be collected. There are already several repositories of MS-derived proteomics data, such as PeptideAtlas (http://www.peptideatlas.org/) and PRIDE (http://www.ebi.ac.uk/pride/).

The PRIDE database offered by the European Bioinformatics Institute is one of the most prominent data repositories of MS-based proteomics data. The main data types stored in PRIDE are protein/peptide identifications, posttranslational modifications, raw mass spectra, and related metadata. As of September 2012, PRIDE contained 25,853 MS-based proteomics experiments, around 11.1 million identified proteins, 61.9 million identified peptides, and 324 million spectra.

5.5 Modeling with different data mining techniques

In this section, we use several different data mining techniques to solve the same protein inference problem. The analysis methods can be either supervised classification or unsupervised cluster analysis. This demonstrates that the same bioinformatics problem can be solved with fundamentally different data mining methods.

5.5.1 A classification approach

The BagReg method [5] formulates the protein inference problem as a standard supervised classification problem, which has three major phases: feature extraction, prediction model construction, and prediction result combination. Figure 5.2 gives an overview of this method.

In *feature extraction*, five features are generated from the original input data for each protein: the number of matched peptides, the number of unique peptides, the number of matched spectra, the maximal score of matched peptides, and the average score of matched peptides. As described in Figure 5.3, these five features are directly

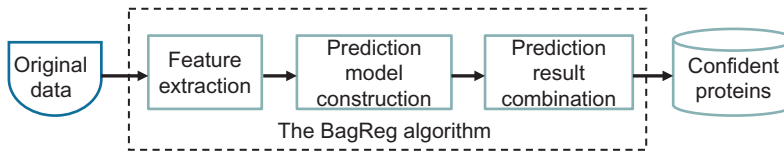
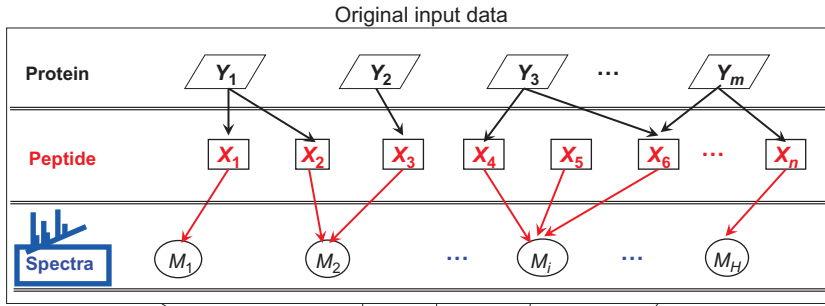


Figure 5.2 An overview of the BagReg method. It is composed of three major steps: feature extraction, prediction model construction, and prediction result combination. In feature extraction, the BagReg method generates five features that are highly correlated with the presence probabilities of proteins. In prediction model construction, five classification models are built and applied to predict the presence probability of proteins, respectively. In prediction result combination, the presence probabilities from different classification models are combined to obtain a consensus probability.



Protein	MP	UP	MS	MSP	AMP
1	90	89	180	1.0	0.92
2	3	1	91	0.98	0.36
3	68	28	12	1.0	0.74
...	...	...	...	...	...
m	25	62	72	0.75	0.99

Extraction results

Figure 5.3 The feature extraction process. Five features are extracted from the original input data for each protein: the number of matched peptides (MP), the number of unique peptides (UP), the number of matched spectra (MS), the maximal score of matched peptides (MSP), and the average score of matched peptides (AMP).

obtained from the input data, and their values are numeric and easy to be calculated. Besides, there is a positive correlation between these feature values and the presence probabilities of proteins in the biological sample. That is, proteins with higher feature values are more likely to be present in the sample than those with lower feature values. This property brings much convenience to the construction of training data set.

In *prediction model construction*, several different learning models are generated independently. Because all five features are positively correlated with the presence probabilities of proteins, the training data set is constructed by taking each of the five features as the class feature. After that, classification methods are exploited to construct a predictive model on the training data and then the classification model is applied to predict presence probabilities for all proteins. Figure 5.4 gives an illustration of a single learning process.

In each single learning process, one feature is selected as the class feature and the other four features are regarded as dependent features. Because there is a positive correlation between the feature value and the protein presence possibility, the set of candidate proteins can be sorted based on the class feature value. Then, a portion of the top-ranked proteins is used as positive training set and a portion of proteins at the end of the sorted list is taken as the negative training set.

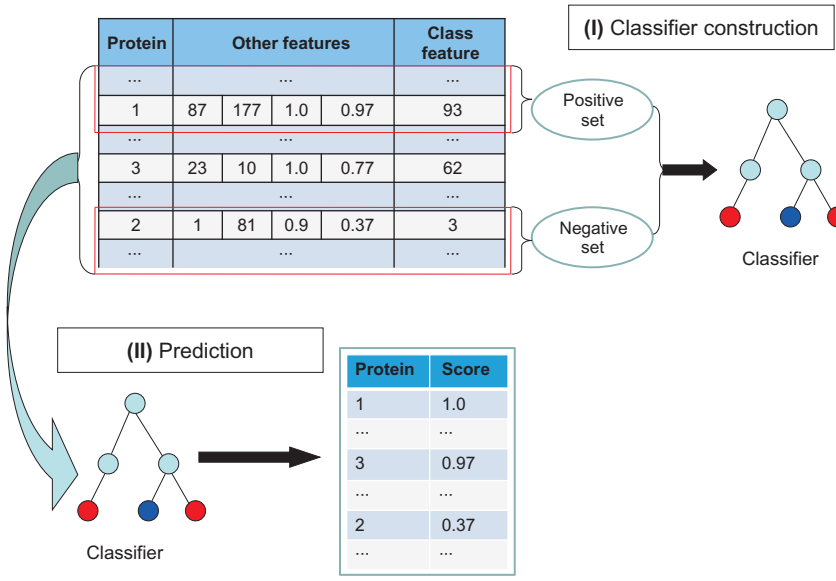


Figure 5.4 A single learning process. Each separate learning process accomplishes a typical supervised learning procedure. The model construction phase involves constructing the training set and learning the classification model. And the prediction phase is to predict the presence probabilities of all candidate proteins with the classifier obtained in the previous phase.

After the training set is generated, a learning model is ready to be built. Any classification model that could produce a probability as the prediction result can be applied. In BagReg, the logistic regression and Bayesian network are used to construct a predictive model on the training set and then predict the presence probabilities for all proteins.

In *prediction result combination*, the five scores of each protein are integrated to obtain a consensus score. The simplest method is to calculate the arithmetic mean of five scores.

5.5.2 A regression approach

The ProteinLasso method [6] formulates the protein inference problem as a constrained Lasso [7] regression problem. First, the probability of each identified peptide can be expressed as the linear combination of protein probabilities, where the coefficients are the conditional peptide probabilities given proteins. Such conditional peptide probabilities are called "peptide detectabilities," which is an intrinsic property of the peptide and can be predicted from existing identification results. Meanwhile, the probability of each identified peptide can be obtained from the peptide identification algorithm and some postprocessing methods.

If we take the protein probabilities as unknown variables and assume that peptide probabilities and peptide detectabilities are known in advance, the protein inference problem can be formulated as a constrained least squares regression problem. Here

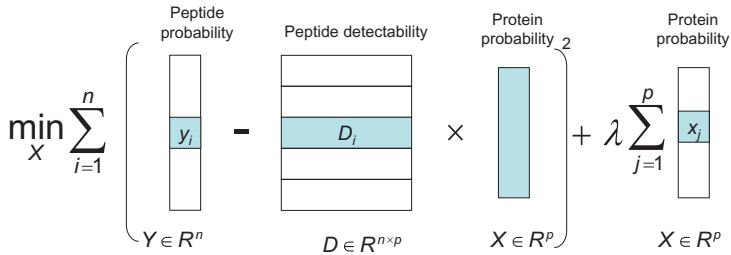


Figure 5.5 The basic idea of ProteinLasso. ProteinLasso formulates the protein inference problem as a minimization problem, where y_i is the peptide probability, D_i represents the vector of peptide detectabilities for the ith peptide, x_j denotes the unknown protein probability of the jth protein, and λ is a user-specified parameter. This optimization problem is the well-known Lasso regression problem in statistics and data mining.

some additional constraints are introduced as the probability of each protein should fall into [0, 1].

Furthermore, an additional penalty term is introduced into the model to make some variables to be zeros since the objective of protein inference is to find a subset of proteins that is truly present in the sample. Such modification leads to a constrained Lasso regression problem, as described in Figure 5.5.

5.5.3 A clustering approach

The protein inference problem can be modeled as a clustering problem as well, see Ref. [8]. Similar to the BagReg method, several different features for each protein can be extracted from the original input data, as shown in Figure 5.3. More precisely, the feature extraction procedure transforms the raw data into a tabular form, in which the columns represent features and rows correspond to proteins.

Although the class label (presence or absence) of each protein is unknown, the fact that these proteins can be divided into two groups is known in advance. One group is the set of proteins that really generates the identified peptides; another group is the set of proteins that is not present. Meanwhile, proteins in the same group have similar feature values. Based on the above observations, the clustering-based approach for protein inference is a two-step procedure: *cluster analysis* and *group identification*.

In the cluster analysis step, existing clustering algorithms such as k-means and hierarchical clustering are applied to partition the set of proteins into two clusters.

In the group identification step, the problem is to select one cluster as the set of truly present proteins. Because each feature is positively correlated with the protein presence probability, the cluster that has larger average feature values is selected.

5.6 Validation: Target-decoy versus decoy-free

How to assess the performance of different protein inference methods is a nontrivial problem. To date, there have already been some proteomics data sets in which the ground truth proteins are known in advance. However, such benchmark data sets

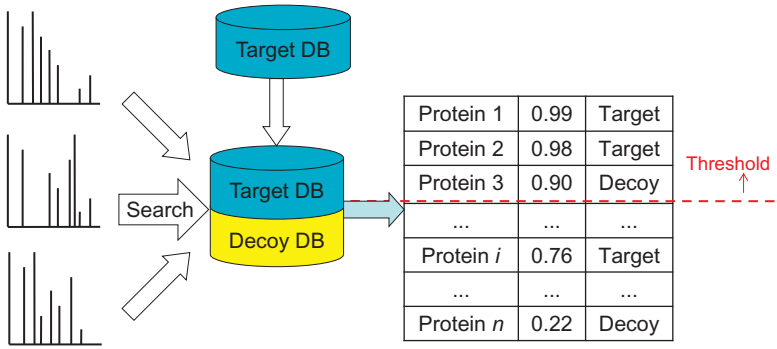


Figure 5.6 The target-decoy strategy for evaluating protein inference results. The MS/MS spectra are searched against the target-decoy database, and the identified proteins are sorted according to their scores or probabilities. The false discovery rate at a threshold can be estimated as the ratio of the number of decoy matches to that of target matches.

usually contain no more than 100 proteins, which cannot reflect the characteristics of real proteomics data sets.

5.6.1 Target-decoy method

The most popular approach for evaluating the identification results in the field of proteomics is the target-decoy strategy. As shown in Figure 5.6, the target-decoy strategy relies on a target-decoy database. This database contains all target protein sequences possibly present in the sample and an equal number of decoy sequences by reversing or reshuffling target protein sequences. During protein identification, the tandem mass spectra are searched against this target-decoy database. To validate the identification results, the false discovery rate (FDR) can be estimated as the ratio of the number of decoy matches to that of target matches.

The target-decoy approach is easy to understand and simple to implement. However, it has some drawbacks. First of all, searching both the target and the decoy database will double the running time of the protein identification procedure. In addition, the FDR estimation result can be unstable, as we usually use only one decoy database of the same size in the identification and evaluation process.

5.6.2 Decoy-free method

Different from the target-decoy approach, the decoy-free evaluation method estimates the FDR without searching the decoy database [9]. The decoy-free method in Ref. [9], as described in Figure 5.7, is based on the null hypothesis that each candidate protein matches an identified peptide totally at random. Under this null hypothesis, it first generates multiple random bipartite graphs with the same structure as the original one, that is, each protein (peptide) is connected to the same number of peptides (proteins).

To generate the null distribution, it is necessary to run the same protein inference method on these simulated graphs to obtain protein scores. However, it is time

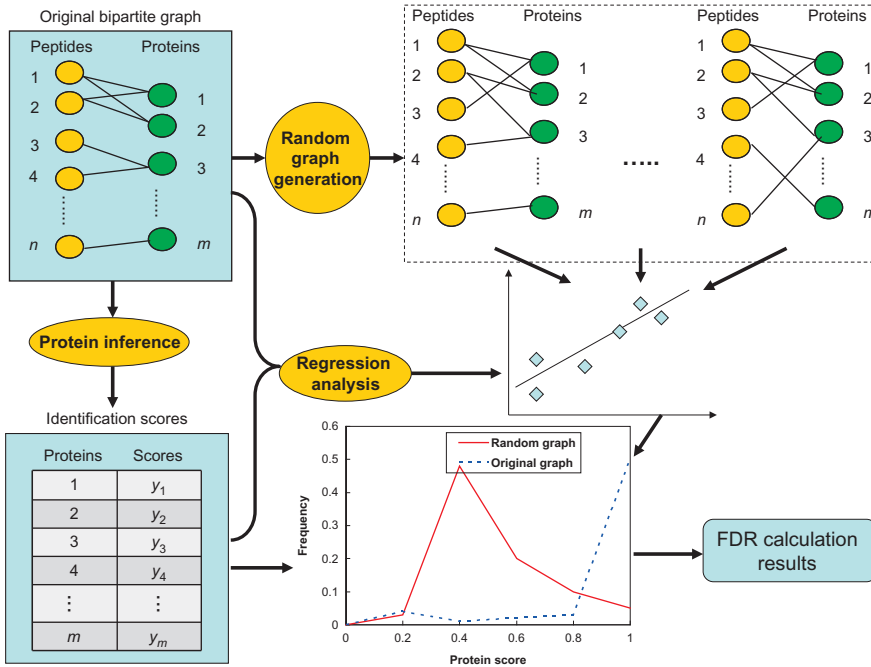


Figure 5.7 An overview of the decoy-free FDR estimation algorithm.

consuming and inconvenient to run some protein inference algorithms for many times. To alleviate this issue, a linear regression model is trained with the original bipartite graph and identification scores given by the target inference algorithm as the input. In this model, the class feature is the protein score, and the dependent features are peptide identification probabilities.

Then, the learned regression model is used as a substitute for the original protein inference method to predict protein scores on randomized graphs. If the null hypothesis that each protein matches an identified peptide by chance is true, then there is no significant difference between the score of each protein in the original graph and those calculated from the random graphs. Therefore, the permutation p-value of one protein can be calculated as the percentage of random graphs that produce a larger score than its score generated on the original graph. Based on these permutation p-values, the FDR at different cut-off thresholds can be derived according to some existing methods (e.g., Ref. [10]).

5.6.3 On unbiased performance evaluation for protein inference

Data mining is a cornerstone of modern bioinformatics. Meanwhile, an unbiased performance evaluation is undoubtedly the cornerstone of data mining research and applications, which provides a clear picture of the strengths and weaknesses of existing approaches.

In the real-world applications of data mining and machine learning methods, there are two closely related and separate problems: model selection and model assessment. In model selection, we estimate the performance of different models to choose the best one. In model assessment or performance evaluation, we test the prediction error of a final model obtained from the model selection process.

The protein inference problem is an instance of prediction task in data mining as well, as shown in Figure 5.8. In model selection, we use the peptide–protein bipartite graph as the input to find a "best" inference model that produces a vector $\bar{Y}$. Each element in $\bar{Y}$ can be either the probability/score that each protein is present or the presence status of each protein (true or false). In model assessment, we compare the predicted vector $\bar{Y}$ with ground truth vector Y to obtain the performance estimates. This is the correct procedure for evaluating and comparing protein inference algorithms.

In contrast, one possible mistake in an incorrect procedure is illustrated at the top of Figure 5.8: the partial or whole ground truth vector Y is used in the model selection process of protein inference algorithms. The problem is that the inference algorithms have an unfair advantage since they "have already seen" the absence/presence information in Y that should only be available during model assessment. In other words, the ground truth information has been leaked to the model selection phase. As a result, the performance estimates of inference algorithms will be overly optimistic. This phenomenon is essentially analogous to the selection bias observed in classification or regression due to feature selection over all samples prior to performance evaluation.

Such biased performance evaluation may occur when we use the target-decoy strategy for comparing the performance of different protein inference methods. In the target-decoy database search and evaluation strategy, a protein is regarded as a true

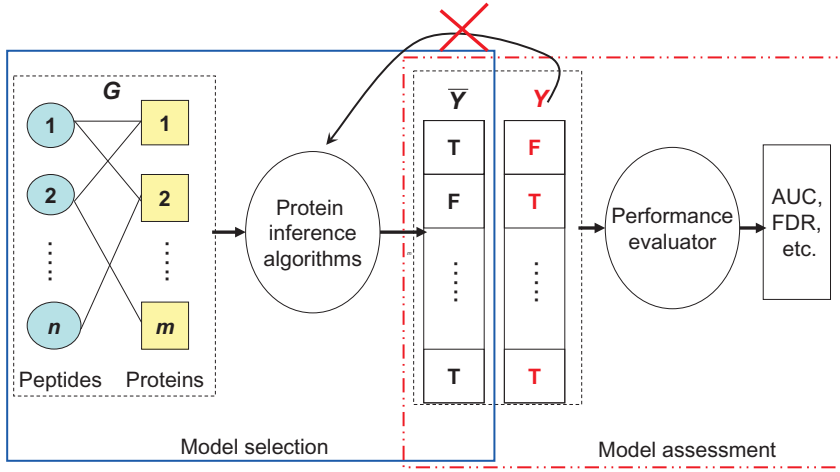


Figure 5.8 The correct and incorrect procedure for assessing the performance of protein inference algorithms. In model selection, we cannot use any ground truth information that should only be visible in the model assessment stage. Otherwise, we may overestimate the actual performance of inference algorithms.

positive if it comes from the target database and as a false positive otherwise. Therefore, the set of target/decoy labels is equivalent to the set of ground truth labels in this context. If we incautiously use the target-decoy information in both the model construction phase and validation phase of one protein inference algorithm, such over-fitting in model selection will lead to an overestimation of its actual performance.

The fact that over-fitting at the level of model selection can have a very substantial deleterious effect in performance evaluation has been widely discussed and recognized in data mining fields and bioinformatics societies. In protein inference, we will face the same problem as well. Therefore, people should be aware of such risk in the future comparison when developing new protein inference algorithms.

5.7 Discussion and future perspective

Protein identification is one of most important problems in shotgun proteomics. Because proteins are more biologically relevant than peptides, it is critical to accurately infer all proteins present in the sample from identified peptides. However, such protein inference problem is still far from being resolved. This is because several technical challenges remain unsolved [4]: the identification coverage problem, the identification ambiguity problem, and the identification validation problem.

Identification coverage problem: Because of the complexity of proteomics data and the limitations of existing peptide identification algorithms, most tandem mass spectra in a typical proteomics experiment cannot be recognized confidently. As a result, only a subset of peptides present in the sample will be identified. This will lead to the exclusion of proteins that have no constituent peptides being identified in the protein inference step. Therefore, some truly present proteins will not be included in the identification results.

Identification ambiguity problem: The ambiguity in protein inference primarily comes from two sources: degenerate peptides (peptides that are shared by more than proteins) and one-hit wonders (proteins that have only one identified peptide). It is generally very difficult to determine which proteins are truly present in the sample if they share the same set of peptides or have only one constituent peptide identified.

Identification validation problem: One of the major problems in computational proteomics is the lack of widely accepted theoretical estimates of statistical significance of protein identifications. The Proteomics Publication Guidelines recommend the use of target-decoy strategy to validate protein identifications. Indeed, if the statistical significance of protein identifications can be estimated accurately in a decoy-free manner, the search in decoy database is not necessary.

In summary, more research efforts still should be devoted to the protein inference problem before the above technical challenges can be solved.

References

[1] A.M. Altelaar, J. Munoz, A.J. Heck, Next-generation proteomics: towards an integrative view of proteome dynamics, Nat. Rev. Genet. 14 (1) (2012) 35–48.

[2] M.S. Kim, et al., A draft map of the human proteome, Nature 509 (7502) (2014) 575–581.
[3] T. Huang, Z. He, A linear programming model for protein inference problem in shotgun proteomics, Bioinformatics 28 (22) (2012) 2956–2962.
[4] T. Huang, J. Wang, W. Yu, Z. He, Protein inference: a review, Brief. Bioinform. 13 (5) (2012) 586–614.
[5] C. Zhao, D. Liu, B. Teng, Z. He, BagReg: protein inference through machine learning, Comput. Biol. Chem. (2015).
[6] T. Huang, H. Gong, C. Yang, Z. He, ProteinLasso: a Lasso regression approach to protein inference problem in shotgun proteomics, Comput. Biol. Chem. 43 (2013) 46–54.
[7] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B (Methodol.) 58 (1) (1996) 267–288.
[8] Y. Zhang, Clustering Algorithm for Mixed Type Data and Its Application, Master's Dissertation, Dalian University of Technology, 2013 (in Chinese).
[9] B. Teng, T. Huang, Z. He, Decoy-free protein-level false discovery rate estimation, Bioinformatics 30 (5) (2014) 675–681.
[10] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. U.S.A. 100 (16) (2003) 9440–9445.