上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

BACHELOR'S THESIS

论文题目：Assessment of Zero-Inflated Model in Proteomics Expression Profile Analysis

学生姓名：吴国经

学生学号：5130809069

专　　业：生物信息

指导教师：李婧

学院(系)：生命科学技术学院

蛋白质组表达谱分析中零膨胀模型效力评估

摘要

鉴定在不同条件下，蛋白质组学的差异表达数据具有重要的科学意义，可以帮助我们找到疾病特异蛋白或蛋白模块，药物靶点，以及加深人类对各类生化反应的认知。近些年的研究已经向我们表明，基于非标记蛋白定量技术的光谱计数定量方法在数据分析层面具有更高的统计效力。但在通常情况下，蛋白光谱计数数据矩阵中会含有大量零值，现有的蛋白质组学差异表达分析方法都还不足以处理如此大量的零值。不过，在生物学其他领域分析当中，已有研究学者使用过零膨胀模型，分析发现零膨胀模型能够处理数据中零值过表达的问题。因此在本研究中，我们提出了基于蛋白互作网络的零膨胀泊松广义线性模型，来帮助我们鉴定蛋白质组学中的差异表达。我们首先进行了仿真模拟来评估零膨胀泊松广义线性模型和其他统计方法间的优劣，随后我们将该方法应用于两套真实案例当中，分别是 CPTAC 的结直肠癌数据和一套非小细胞肺腺癌数据。在评估比较中我们最终发现，零膨胀泊松广义线性模型能鉴定出更多统计显著的疾病相关差异表达蛋白模块。

关键词：差异表达鉴定，零膨胀模型，广义线性回归，光谱计数，蛋白网络，结直肠癌

Assessment of Zero-Inflated Model in Proteomics Expression Profile Analysis

ABSTRACT

Identify proteomics differential expression (PDE) among distinct conditions plays an important role in finding disease-specific proteins and modules, drug targets as well as understanding the mechanism of biochemical reactions. Recent studies have demonstrated that label-free quantification by spectral counting have an increase in statistical power of data analysis. But often, plenty of cells in these spectral counting protein matrices have zero counts, existing methods in PDE analysis are not capable of handling the excess number of zeros well. However, a zero-inflated (ZI) distribution model has been applied to other biological filed which proved itself able to fit the overexpression of zeros. In this study, we proposed a network module-based zero-inflated generalized linear model (ZIGLM) for identify PDE of spectral count data. We first conducted simulation studies to compare efficacy between ZIGLM and other statistical methods. Then, we applied our method to CPTAC colorectal cancer data set and a nonsmall cell lung cancer data set. In comparison with other distribution models or common PDE methods, ZIGLM could identify more significantly biological modules that are highly related to cancer.

Key words: differential expression analysis, zero-inflated model, generalized linear regression, spectral count, biological network module, colorectal cancer.

Contents

Chapter One Introduction

High-throughput shotgun proteomics is now causing a great impact on biological research due to the emergence of liquid chromatography tandem mass spectrometry (LC-MS/MS). Among all the MS-driven protein quantification methods, label-free method has proved to be more robust and less complicate than tagged or isotope labelled methods ([1], J. Cox and M. Mann, 2011: 273-99, 2], S. R. Langley and M. Mayr, 2015: 83-92), e.g. isobaric tags for relative and absolute quantitation (iTRAQ) ([3], P. L. Ross, 2004: 1154-69), tandem mass tags (TMT) ([4], L. Dayon and J. C. Sanchez, 2012: 115-27), and stable isotope labelling by amino acids in cell culture (SILAC) ([5], M. Mann, 2006: 952-8). Label-free quantitative approach either employs ion intensity changes from peptide peak areas or heights ([6], H. Wang, 2012: 487-501), or is based on number of peptide spectrum matches (PSMs; spectral counts) from a protein to supersede protein abundance ([1], J. Cox and M. Mann, 2011: 273-99). Recent studies have illustrated that spectral count can be more sensitive than ion intensity while possess high technical reproducibility ([7], B. Zhang, 2006: 2909-18, 8], W. M. Old, 2005: 1487-502), hence, more accurate and comprehensive understanding of biology can be penetrated in light of researches using spectral count data.

1.1 Shotgun Proteomics

The word proteomics denotes the comprehensive expressed proteins in a specific state of an organism or cell line ([1], J. Cox and M. Mann, 2011: 273-99). Its main purpose is to obtain a global view at the protein level on the analogy of what has been done at the level of DNA and RNA.

However, despite the similarities between proteomics and genomics, there exists many discrepancies. Genomics focus on genotype while proteomics focus on phenotype which is determined by both the genotype and the surrounding environment. Besides, the technology between these two fields are different, genomics takes advantage of scalable and refined technologies for sequencing, whereas proteomics employed some novel techniques, like mass spectrometric (MS).

Apart from structural proteomics, the MS-based proteomics has three ramifications: expression proteomics, modification proteomics and interaction proteomics. The main purpose of expression proteomics is to quantify the absolute or relative amount of proteins in a sample. It has some advantages than transcriptomics because it takes the regulation of posttranscriptional into account. In this paper, we majored in expression proteomics.

1.2 Differential Expression Protein Analysis

Proteomics differential expression (PDE) analysis plays a key role in comparative proteomics, these includes significant protein analysis, clustering and classification ([9], A. Pandini, 2013: 642-51, 10], C. Pizzuti and S. E. Rombo, 2014: 1343-52), and network analysis ([11], U. Kirik, 2012: 2955-67, 12], L. Liu and J. Ruan, 2013: 218-21, 13], B. Chen, 2014: 177-94). Heretofore, several methods have been proposed to detect significant proteins under two distinct conditions based on spectral count data. Table 1 shows some representative examples of them since 2007 ([2], S. R. Langley and M. Mayr, 2015: 83-92, 14], H. Choi, 2008: 2373-85, 15], X. Fu, 2008: 845-54, 16], K. M. Little, 2010: 1212-22, 17], M. C. Leitch, 2012: 89-98, 18], O. E. Branson and M. A. Freitas, 2016: 23-32). Besides, Zhang et al. proposed a generalized G-test to detect differential expression proteins under multiple experimental conditions ([7], B. Zhang, 2006: 2909-18).

Table 1.1 Previous Studies on Detecting Differential Expression Proteins Using Label-Free Spectral Count Data

article	methods	compared with
Xiaoyun Fu, et al. 2007	SpI[a]	student's t-test, G-test, Bayesian t-test, SAM[b]
Hyungwon Choi, et al. 2008	QSpec	PLGEM-StN[c]
Kristina M. Little, et al. 2010	ReSASC[d]	LPE[e] test, Ranks Products
Matthew C. Leitch, et al. 2012	NB[f] model	QSpec, quasi-Poisson model
Sarah R. Langley, et al. 2015	-	t-test, SAM, NSAF, NSAF-PLGEM, SpI, DESeq, QSpec
Owen E. Branson, et al. 2016	MultiSpec	-

[a]SpI, spectral index.

[b]SAM, significance analysis of microarray data.

[c]PLGME, power law global error model.

[d]ReSASC, resampling-based significance analysis for spectral counts.

[e]LPE, local-pooled error.

[f]NB, negative binomial

1.3 Module-based Differential Expression Analysis

The main problem for differential protein analysis is that it doesn't take the information of protein-protein interactions (PPI) into account. As we know, most biological functions are implemented by sets of interrelated proteins, and sometimes, the statistical significance analysis of abundance change in individual proteins may fail, because the differences are modest relative to the noises in shotgun proteomics.

So there is an urgent need for a high level statistical significance analysis that can utilize the protein-protein interaction networks while maintain high efficacy.

In previous, Xu et al. proposed a module-based generalized linear model to detect differential expression modules instead of proteins ([19], J. Xu, 2014: 5743-50). They used a negative binomial model and made use of both KEGG pathways and PPI modules. The results showed that by taking PPI information into account, this method can increase the statistical power and identify some important pathways or modules that has latent value.

1.4 Objectives of This Study

Despite all the studies have been done, there remains a thorny problem still far from been settled: the over expression of zero in spectral counts, or so-called, zero inflation (ZI). Spectral counts data usually contain over 20% of zero counts ([20], B. Zhang, 2014: 382-7, 21], N. Pavelka, 2008: 631-44, 22], T. Kikuchi, 2012: 916-32, 23], C. Shao, 2010: 313-26), some of them are “true zeros” indicating that the protein didn’t express in the corresponding sample, while other zero counts called “false zeros” are simply modeled zero counts which indicating the protein hadn’t been detected yet ([24], L. Huang, 2017: 471-88). There are a bunch of reasons that can account for detection failure, such as MS instrumental error ([25], F. Gosetti, 2010: 3929-37, 26], T. M. Annesley, 2003: 1041-4, 27], P. J. Taylor, 2005: 328-34), protein degradation during sample treatment, or incomplete reference database for search engine ([28], J. Cox and M. Mann, 2008: 1367-72, 29], T. Muth, 2010: 1522-4). Consequently, ZI will cause the distribution highly skewed, and therefore, hypothesis testing under normal distribution assumption or statistical models that cannot cope with the overdispersion are not suitable for these data. Extant methods choose either ignore the ZI effect or delete low abundance proteins during sample preprocessing, but such a

large number of zero counts will no doubt contain considerable information. Based on that, it's necessary to come up with an approach that can deal with ZI while attain high sensitivity and specificity.

Previously, a zero-inflated model has been proposed and used for regression analysis. Researchers had employed the zero-inflated model in mapping species abundance ([30], O. Lyashevska, 2016: 532-43), or drug safety signal detection ([24], L. Huang, 2017: 471-88). The results showed that zero-inflated model can fit the ZI problem and increased statistical power than other methods.

In this study, we proposed a module-based zero-inflated generalized linear model (ZIGLM) for analyzing differential expression modules in shotgun proteomics using spectral count data. We first used simulation datasets to investigate the ability of zero-inflated models and other models with respect to several treatment measures: 1) effect size fc, or abundance fold change between two distinct conditions, 2) proportion size p, or the percentage of proteins within a module that are differentially expressed and 3) zero ratio "π", or the percentage of true zeros within a dataset. Then, we applied our ZI model to two cancer datasets to identify significant differentially expressed modules.

Chapter Two Materials and Methods

2.1 Proteome Data Collection

We used three publicly available label-free shotgun proteomics data sets. No reanalysis of the raw MS file was conducted on these data sets, which means the spectral count data matrices were directly taken from supplemental materials offered by these publications. The first is hibernating arctic ground squirrels data set, it contains a total of 3594 proteins among 12 sample. All proteins were required to have at least two unique peptides hits. 12 samples were obtained from three different stages: two of them are from hibernating stages between torpor and arousal: late torpor (LT) and early arousal (EA), and the third is called postreproduction (PR) as non-hibernating control, each stage has 4 samples ([23], C. Shao, 2010: 313-26).

The second data set was from CPTAC colorectal cancer (CRC) study, it contains a total of 7244 proteins among 120 samples ([20], B. Zhang, 2014: 382-7). The spectral count data were summarized at gene group level, and for each gene group, gene with the shortest protein length was chosen to represent the group. 90 of the whole samples are tumor samples (5 of them have duplicated samples) and the rest are normal samples. While removing duplicates, the sample with a larger total spectral count was remained, then a quantile normalization was performed on it

And the third was from a nonsmall cell lung cancer study, it contains a total of 5123 proteins among 16 samples. Proteins were summarized at protein group level, if several proteins mapped to the same gene symbol, we randomly chose one to represent the group ([31], C. The UniProt, 2017: D158-D69). Among the 16 samples, 4 of them were from stage I adenocarcinomas (ADC), 4 of them were from stage I

squamous cell carcinomas (SCC) and 8 samples from normal condition ([22], T. Kikuchi, 2012: 916-32).

2.2 Protein Network and Module

For integrated biological network, we chose KEGG pathway database ([32], M. Kanehisa, 2017: D353-D61, 33], M. Kanehisa and S. Goto, 2000: 27-30). The human KEGG pathway was downloaded from GSEA on April 6th, 2017 ([34], V. K. Mootha, 2003: 267-73). It contains a total of 186 pathways, every KEGG pathways were treated as modules.

2.3 Zero-Inflated Model

Zero-inflated model is a model class capable of dealing with ZI.([35], A. Zeileis, 2008: 1-25) It's a mixture model composed of a point mass at zero $M_{([36],\ I.\ Alroy\ and\ Y.\ Yarden,\ 83\text{-}6)}(y)$ and a count distribution $f_{count}(y;\ x,\ \theta_1)$. For the first part of zero-inflated model, a binary model such as Bernoulli is used for modeling the hidden state (expressed vs. unexpressed, or true zero vs. others). The second part is used for modeling spectral counts distribution, we assume it follows Poisson, negative binomial or geometric distribution. And this is so-called, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. Since the geometric distribution is a special case of negative binomial, it doesn't have an abbreviation like ZIP or ZINB. The general probability density function (PDF) formula is given below:

$$f_{ZI}(y;\ x,\ \theta_1,\ z,\ \theta_2) = f_{zero}(0;\ z,\ \theta_2) \cdot M_{\{0\}}(y) + \left(1 - f_{zero}(0;\ z,\ \theta_2)\right) \cdot f_{count}(y;\ x,\ \theta_1)$$

(2-1)

Where $\theta_1=(\theta_{11}, \theta_{12}, \cdots, \theta_{1k})$, $\theta_2=(\theta_{21}, \theta_{22}, \cdots, \theta_{2k})$, and the probability of observing a true zero is given by $\pi = f_{zero}(0; z, \theta_2)$. This model is compiled in *psyl* package in R ([37], H. Wickham, 2011: 1-29).

2.4 Generalized Linear Model

Generalized linear regression model, or generalized linear model (GLM) is a flexible generalization of ordinary linear model (OLM) that allows the response variables to have error distribution models other than normal distribution ([38], G. H. Dunteman and M.-H. R. Ho, 2006: x, 72 p.). Also, instead of linear relation, GLM generalize OLM by allowing the linear model to be related with response variables via a link function. In our study, we proposed a network module-based model for the PDE analysis based on previous study:

$$log\left(y_{ij}\right) = \mu_{ij} + G_{ij} + M_{ij} + G_{ij} * M_{ij} \quad (2\text{-}2)$$

Here, $y_{i,j}$ represents the expected spectral counts of protein i in sample j. $\mu_{i,j}$ is the overall mean. $G_{i,j}$ and $M_{i,j}$ are indicator variables. In a case-control study, $G_{i,j} = 1$ means the cell belongs to case group and $G_{i,j} = 0$ means the cell belongs to control group. Similarly, $M_{i,j} = 1$ means protein i belongs to the specified module and $M_{i,j} = 0$ means not. And $G_{i,j} * M_{i,j}$ is an interaction term. By taking protein interactions into account, this model is supposed to identify statistically significant modules.

2.5 Data Simulation

Simulation data sets are synthesized based on the hibernating arctic ground squirrels data set. The original data set contains 3594 (n) proteins and 12 (m) samples, and we shuffled spectral counts across the proteins for five times to eliminate the effect caused by different sample ([14], H. Choi, 2008: 2373-85). Then we simulate data with the same number of proteins and samples by using following steps (Figure 1).

First, we determine the true zero cells. The percentage of true zero cells in sample j is assumed to be π. We set four levels for it: 0%, 30%, 50% and 70%. True zero cells are simulated using Bernoulli probability model:

$$z_{ij} \sim Bernoulli(\frac{\pi n}{P_{0j}}) \quad i=1,\ldots, n \tag{2-3}$$

Where P_{0j} is the total number of observed zero cells in sample j. If $z_{i,j} = 1$, it means $y_{i,j}$ is true zero cells.

Second, all 3594 proteins are randomly assigned to 100 (M) modules, every modules' size is determined by using uniform distribution:

$$s_t = \frac{q_t}{\sum_{k=1}^{M} q_k} * n \qquad q_t \sim U(a, b) \qquad t=1,\ldots,M \tag{2-4}$$

Here we set a = 0.1 and b = 1 in case of empty module occurs. q_t is an indicator, s_t represents the module size. We choose the first module as signal module, and the total number of $s_t * p$ proteins were randomly chosen from the first module. We designate these proteins as signal proteins with effect treatment added to the case group. We set three levels for p: 30%, 50% and 80%.

Finally, we simulate spectral counts matrix by using multinomial model. Under the null hypothesis, all proteins' abundance should have no significant difference between samples, then the distribution of spectral counts will be like this:

$$\left(y_{i1}, \ldots, y_{ij}, \ldots, y_{im}\right) | X_i \sim Multinomial(X_i, (\frac{Y_1}{N}, \ldots, \frac{Y_j}{N}, \ldots, \frac{Y_m}{N})) \quad (2\text{-}5)$$

Where X_i is the sum of spectral counts of protein i, Y_j is the sum of spectral counts of sample j, and N is the total sum of spectral counts of all proteins in all samples. The constraint is that $0 \leq \frac{Y_1}{N} \leq 1$ and $\sum_{j=1}^{m} \frac{Y_j}{N} = 1$. And under the alternative hypothesis, the data are simulated using following equation:

$$\left(y_{i1}, \ldots, y_{im}\right) | X_i \sim Multinomial(X_i, (rr_{i,1} \times r_i \times \frac{Y_1}{N}, \ldots, rr_{i,m} \times r_i \times \frac{Y_m}{N})) \quad (2\text{-}6)$$

Where $rr_{i,j}$ is the relative reporting rate for protein i in sample j, and r_i is interpreted as the baseline rate of protein i. Similarly, the constraint is $0 \leq rr_{i,j} \times r_i \times \frac{Y_j}{N} \leq 1$ and $\sum_{j=1}^{m} rr_{i,j} \times r_i \times \frac{Y_j}{N} = 1$. If $z_{i,j} = 1$, then we assign $rr_{i,j} = 0$. And for cells with $z_{i,j} = 0$, if the corresponding protein is designated as signal protein (protein with effect treatment), we assign $rr_{i,j}$ in case group equals to fc, and $rr_{i,j} = 1$ for control group. For all other cases not mentioned above, we assign $rr_{i,j} = 1$. We set five levels for fc: 1.2, 1.5, 2.0, 3.0 and 4.0 (more details see appendix).

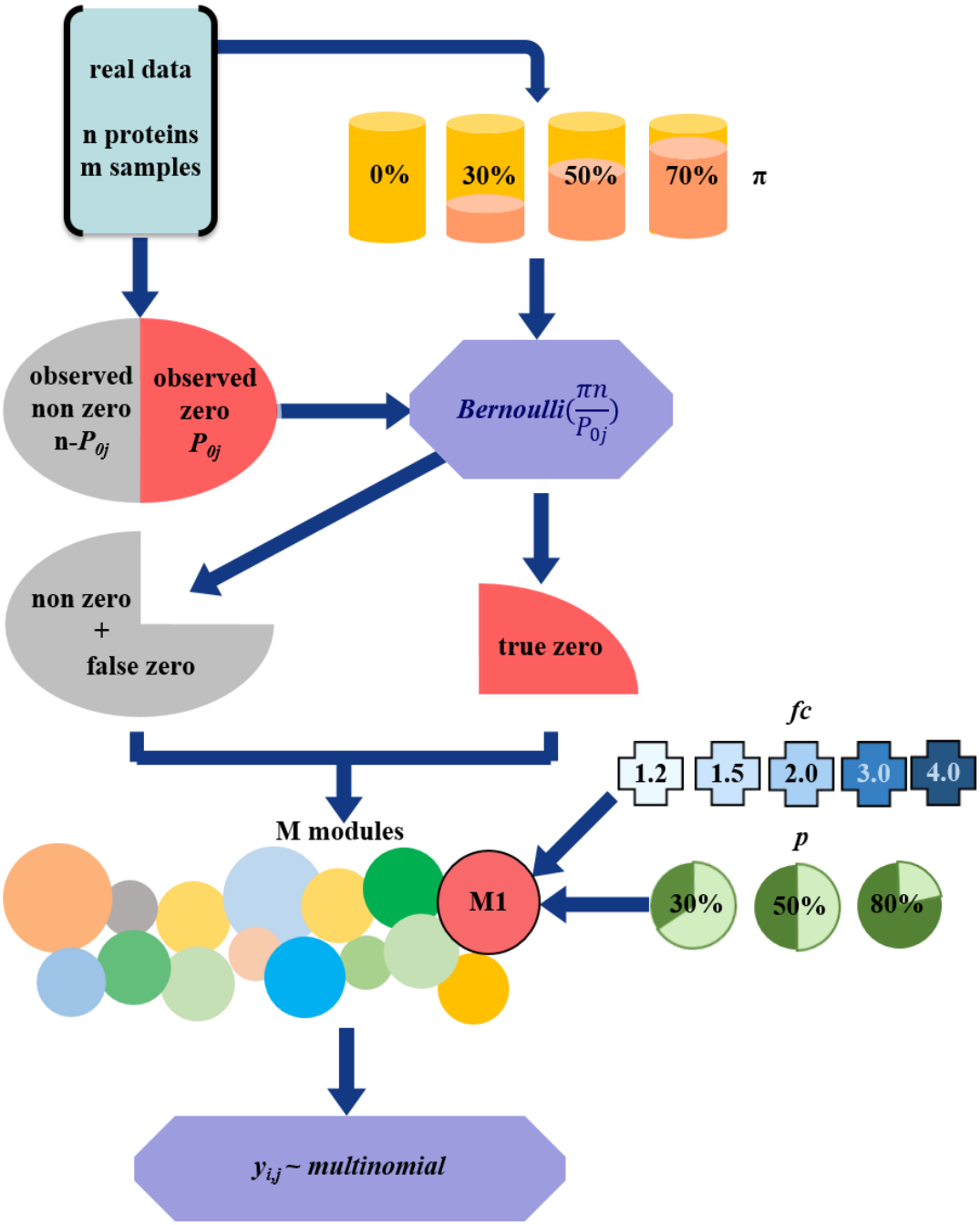

Figure 2.1 Basic data simulation procedure. Three effect treatments π (true zero ratio), p (percentage of signal proteins in the first module, M1) and fc (fold change) are plot.

Chapter Three Results and Discussion

3.1 Simulation Study

In our simulation study, we assessed only ZIP model (because ZINB model has a lower performance than ZIP based on our study) in three different treatment measures: π, p and fc, and compared it with NB model. For each mixed π, p and fc, we synthesized 1000 simulation data sets and obtained the p-value for every module, then calculated AUC (Area Under the receiver operating characteristic Curve) for every data sets results (Figure 2).

From the two tables in Figure 1, we can see that: 1) for each fixed π, p and fc, ZIP model has a better performance than NB, 2) when data set contains small percentage of true zero, and signal proteins have high fold change between case and control group, ZIP and NB are tend to have similar performance and 3) ZIP shows high performance when data set has high π and low fold change in signal proteins. These demonstrate that ZIP model is capable of handling ZI while attain high sensitivity to detect subtle differential expression. We also implemented Wilcoxon rank-sum test followed by GSEA (Wilcoxon-GSEA) on these simulation data sets ([39], X. Wang, 2011: 879-80, 40], A. Subramanian, 2005: 15545-50), but the Wilcoxon rank-sum test failed to detect any differential expression protein firstly. It's because the percentage of signal proteins ($s_t * p / P$) is so small that there will be no significant signals after multiple hypotheses testing adjustment ([41], Y. Benjamini, 2001: 279-84), which indirectly proved that the module-based ZIGLM has high sensitivity.

AUC for NB model of 1000 simulation study

p \ fc	1.2	1.5	2	3	4	π
0.3	0.500	0.501	0.555	0.721	0.804	0
	0.500	0.500	0.536	0.687	0.793	0.3
	0.502	0.502	0.511	0.648	0.723	0.5
	0.500	0.505	0.511	0.602	0.677	0.7
0.5	0.500	0.509	0.700	0.865	0.914	0
	0.500	0.505	0.673	0.851	0.907	0.3
	0.502	0.506	0.621	0.802	0.850	0.5
	0.512	0.524	0.575	0.750	0.808	0.7
0.8	0.500	0.637	0.858	0.935	0.957	0
	0.500	0.603	0.854	0.926	0.954	0.3
	0.504	0.558	0.796	0.899	0.917	0.5
	0.526	0.573	0.739	0.845	0.882	0.7

0.5
0.6
0.7
0.8
0.9
1

AUC for ZIP model of 1000 simulation study

p \ fc	1.2	1.5	2	3	4	π
0.3	0.587	0.664	0.812	0.923	0.955	0
	0.696	0.690	0.792	0.910	0.935	0.3
	0.781	0.725	0.740	0.824	0.839	0.5
	0.883	0.842	0.807	0.830	0.860	0.7
0.5	0.614	0.741	0.889	0.969	0.986	0
	0.780	0.696	0.860	0.953	0.964	0.3
	0.841	0.757	0.747	0.852	0.890	0.5
	0.920	0.900	0.830	0.840	0.901	0.7
0.8	0.658	0.807	0.955	0.985	0.994	0
	0.853	0.697	0.893	0.975	0.983	0.3
	0.884	0.790	0.738	0.871	0.934	0.5
	0.939	0.906	0.838	0.834	0.906	0.7

Figure 3.1 AUC of 1000 simulations for ZIP and NB model under different parameters. π = true zero ratio, p = percentage of signal proteins in first module, and fc = fold change. The color in each cell is correspond with AUC value.

3.2 Real Data Application

We applied our methods on the CPTAC colorectal cancer data set to help us identify differentially expressed pathways, which may be potentially related to colorectal

cancer. And the nonsmall cell lung cancer data set was used as a supplement. We still performed NB, ZIP and Wilcoxon-GSEA comparatively to see how these three methods implemented. We used the false discovery rate (FDR) adjusted p-value as indicator, and the threshold was set at 0.01.

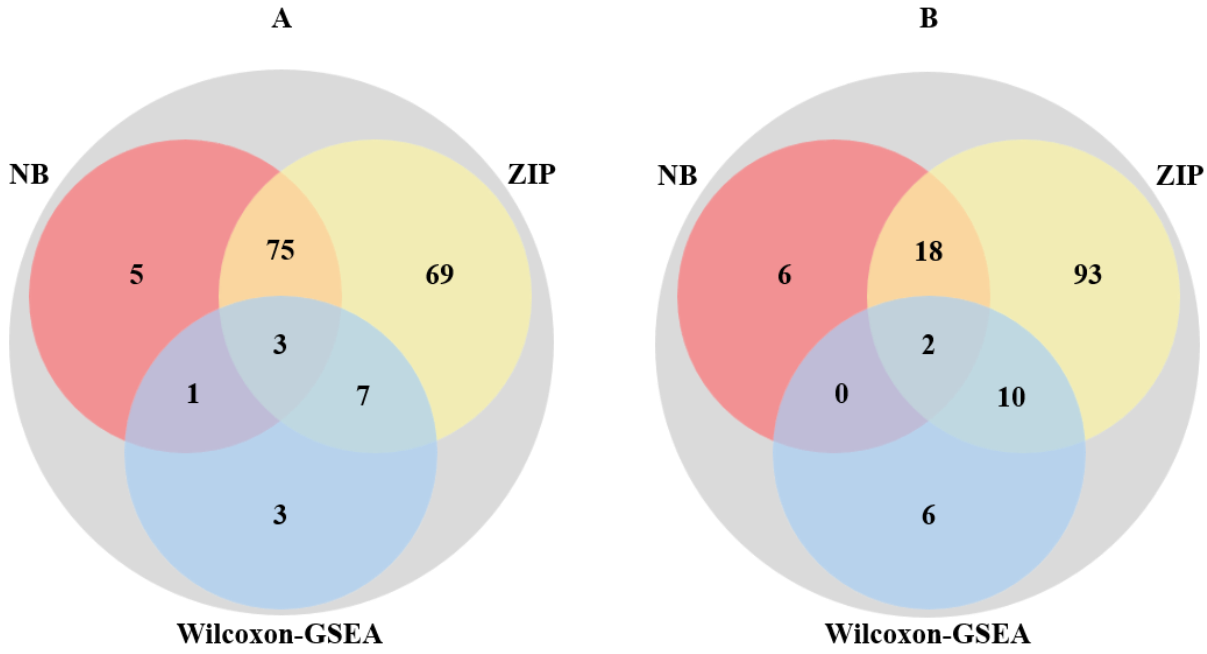

Figure 3.2 Venn diagrams of three methods results: NB, ZIP and Wilcoxon-GSEA. Red circle represents NB, yellow circles represents ZIP and blue circle represents Wilcoxon-GSEA method. (A) Results for colorectal cancer data. (B) Results for nonsmall cell lung cancer data.

The Venn diagrams (Figure 3) show that ZIP can identify the most differentially expressed pathways and its result can cover most of others results (cover 92.86% of NB, 71.43% of Wilcoxon-GSEA). But we couldn't help to notice that the size of ZIP's results was way too big, so we dived into the results to inspect the details. We focused on the unique result produced by ZIP, in other words, pathways that regarded as differentially expressed only by ZIP. Table 2 lists the top 10 pathways (ranked by adjusted p-value) in unique ZIP result for CRC data, some of them have been shown to be associated with CRC. Take the extracellular matrix (ECM) receptor interaction pathway as example, ECM-receptors are a complex mixture of transmembrane

molecules, they mediate interactions between cells and the ECM. These interactions lead to a direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis ([42], W. Hollas, 1991: 3690-5, 43], W. G. Stetler-Stevenson, 1993: 541-73), which also play an important role in colorectal cell carcinogenesis. Other pathways, like focal adhesion ([44], L. V. Owens, 1995: 2752-5, 45], W. G. Cance, 2000: 2417-23), glycolysis gluconeogenesis,([46], X. Bi, 2006: 1119-30) nicotinate and nicotinamide metabolism ([47], L. S. Chen, 2010: 860-71), regulation of actin cytoskeleton ([48], M. Bienz and H. Clevers, 2000: 311-20, 49], P. Gulhati, 2011: 3246-56), and adherens junction have also been proved to be associated with CRC ([48], M. Bienz and H. Clevers, 2000: 311-20, 50], S. T. Mees, 2009: 361-8).

Table 3.1 Top 10 Pathways in Unique ZIP Result for Colorectal Cancer Data

Pathway	adjusted p-value	# of match	pathway size
ECM[a] receptor interaction	0	63	84
hypertrophic cardiomyopathy (HCM)	0	45	85
arrhythmogenic right ventricular cardiomyopathy	0	46	76
focal adhesion	4.35E-301	142	201
glycolysis gluconeogenesis	1.70E-284	51	62
nicotinate and nicotinamide metabolism	6.23E-249	19	24
regulation of actin cytoskeleton	2.17E-248	137	216
adherens junction	8.61E-177	55	75
protein export	7.22E-156	20	24
starch and sucrose metabolism	9.23E-147	31	52

[a]ECM, extracellular matrix.

We also extracted all the signaling pathways in unique ZIP results for CRC data (Table 3), and there appears to have a large number of cancer-related signaling pathways. Some are the most well-known signaling pathways, like MAPK signaling pathway ([51], R. B. Corcoran, 2012: 227-35, 52], J. Lascorz, 2010: 1612-9, 53], J. Y. Fang and B. C. Richardson, 2005: 322-7), Wnt signaling pathway ([48], M. Bienz and H. Clevers, 2000: 311-20, 54], H. Suzuki, 2004: 417-22), TGF beta signaling pathway, ErbB signaling pathway ([36], I. Alroy and Y. Yarden, 1997: 83-6, 55], R. A. Gupta and R. N. Dubois, 2001: 11-21), p53 signaling pathway ([56], N. R. Rodrigues, 1990: 7555-9, 57], J. Vanamala, 2010: 238), and mTOR signaling pathway ([49], P. Gulhati, 2011: 3246-56, 58], S. M. Johnson, 2010: 767-76, 76-8, 59], F. V. Din, 2012: 1504-15 e3), while others take an important part in immune system or signal transduction. Here we took out the exceptional insulin signaling pathway to try to find a reasonable interpretation (Figure 4). We calculated the log transformed mean spectral count fold change between case and control group, and colored these nodes correspondingly. The map shows that this insulin signaling pathway has an upstream/downstream connection with MAPK signaling pathway, apoptosis and cell proliferation and differentiation. The insulin pathway also include some cancer-related nodes, like P13K, mTOR, Ras, Raf, etc ([49], P. Gulhati, 2011: 3246-56, 58], S. M. Johnson, 2010: 767-76, 76-8, 60], C. S. Karapetis, 2008: 1757-65, 61], S. Khambata-Ford, 2007: 3230-7, 62], S. Benvenuti, 2007: 2643-8, 63], M. Sugiyama, 2009: 339-44). The whole results from nonsmall cell lung cancer data are analogous to the results from CRC data, so no more details are narrated for it.

From the result, we can see that the module-based ZIGLM method is more sensitive than other methods and can capture the subtle change between two groups, which can promote our understanding on pathology of complex cancers.

Table 3.2 14 Signaling Pathways in Unique ZIP Result for CRC Data

Pathway	adjusted p-value	# of match	pathway size	KEGG class
MAPK signaling pathway	5.94E-65	122	267	Signal transduction
Fc epsilon RI signaling pathway	4.70E-28	46	79	Signal transduction
VEGF signaling pathway	6.56E-26	41	76	Signal transduction
Wnt signaling pathway	2.95E-23	70	151	Signal transduction
TGF beta signaling pathway	1.71E-10	32	86	Signal transduction
insulin signaling pathway	5.24E-09	86	137	Endocrine system
B cell receptor signaling pathway	2.92E-07	51	75	Immune system
ErbB signaling pathway	3.46E-07	52	87	Signal transduction
p53 signaling pathway	1.81E-05	35	69	Signal transduction
mTOR signaling pathway	0.000128477	26	52	Signal transduction
T cell receptor signaling pathway	0.000216691	61	108	Immune system
notch signaling pathway	0.000446209	18	47	Signal transduction
RIG-I-like receptor signaling pathway	0.001227643	33	71	Immune system
chemokine signaling pathway	0.001970305	90	190	Signal transduction

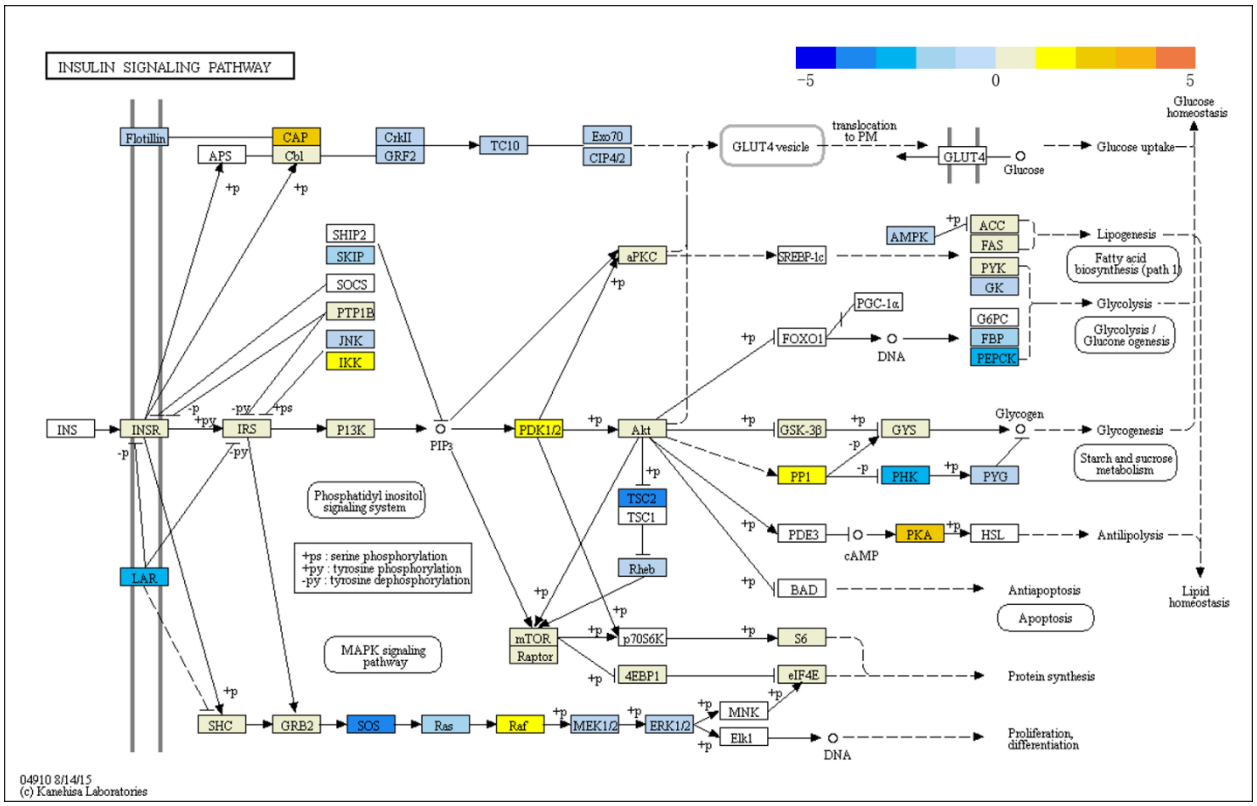

Figure 3.3 Map of insulin signaling pathway. These nodes are colored corresponding to the fold change of protein expression between case and control group.

Chapter Four Conclusions and Future Plan

In summary, we have proposed a module-based zero-inflated generalized linear model. This model can simulate the zero inflation effect and extract extra information from data sets, it also takes the protein-protein interaction network into account which allows us to identify differentially expressed modules. Then we implemented simulation study to assess this method's efficacy in three different effect measurement: π, p and fc. By comparing with negative binomial model, the simulation study demonstrates that ZIGLM possesses a good efficacy than NB, and performs well especially when data set contains too manty zero cells. Finally, we applied our model to two cancer data sets, and compared the results between ZIP, NB and Wilcoxon-GSEA. These two results are in conformity with each other, which show that our method can capture more cancer-related differentially expressed pathways. And these pathways can then function in many fields like disease diagnosis and cancer drug development.

However, in this article, we only discussed the top 10 pathways and all signaling pathways from CRC data set result, other pathways may need further validation and more informational conclusions can be drawn from it. Also, we only applied this method to module level statistical analysis, it may increase the efficacy of individual protein analysis but that requires further validation. And the effect of sample size was not assessed either. Finally, this method together with the whole analysis process can be integrated in R package and make the statistical analysis of expression proteomics more convenient.

REFERENCE

[1] COX J, MANN M. Quantitative, high-resolution proteomics for data-driven systems biology [J]. Annual review of biochemistry, 2011, 80(273-99.
[2] LANGLEY S R, MAYR M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics [J]. Journal of proteomics, 2015, 129(83-92.
[3] ROSS P L, HUANG Y N, MARCHESE J N, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents [J]. Molecular & cellular proteomics : MCP, 2004, 3(12): 1154-69.
[4] DAYON L, SANCHEZ J C. Relative protein quantification by MS/MS using the tandem mass tag technology [J]. Methods in molecular biology, 2012, 893(115-27.
[5] MANN M. Functional and quantitative proteomics using SILAC [J]. Nature reviews Molecular cell biology, 2006, 7(12): 952-8.
[6] WANG H, ALVAREZ S, HICKS L M. Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two Chlamydomonas reinhardtii strains of interest for biofuels engineering [J]. Journal of proteome research, 2012, 11(1): 487-501.
[7] ZHANG B, VERBERKMOES N C, LANGSTON M A, et al. Detecting differential and correlated protein expression in label-free shotgun proteomics [J]. Journal of proteome research, 2006, 5(11): 2909-18.
[8] OLD W M, MEYER-ARENDT K, AVELINE-WOLF L, et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics [J]. Molecular & cellular proteomics : MCP, 2005, 4(10): 1487-502.
[9] PANDINI A, FRACCALVIERI D, BONATI L. Artificial neural networks for efficient clustering of conformational ensembles and their potential for medicinal chemistry [J]. Current topics in medicinal chemistry, 2013, 13(5): 642-51.
[10]PIZZUTI C, ROMBO S E. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods [J]. Bioinformatics, 2014, 30(10): 1343-52.
[11]KIRIK U, CIFANI P, ALBREKT A S, et al. Multimodel pathway enrichment methods for functional evaluation of expression regulation [J]. Journal of proteome research, 2012, 11(5): 2955-67.
[12]LIU L, RUAN J. Network-based Pathway Enrichment Analysis [J]. Proceedings IEEE International Conference on Bioinformatics and Biomedicine, 2013, 218-21.
[13]CHEN B, FAN W, LIU J, et al. Identifying protein complexes and functional modules--from static PPI networks to dynamic PPI networks [J]. Briefings in bioinformatics, 2014, 15(2): 177-94.

[14]CHOI H, FERMIN D, NESVIZHSKII A I. Significance analysis of spectral count data in label-free shotgun proteomics [J]. Molecular & cellular proteomics : MCP, 2008, 7(12): 2373-85.
[15]FU X, GHARIB S A, GREEN P S, et al. Spectral index for assessment of differential protein expression in shotgun proteomics [J]. Journal of proteome research, 2008, 7(3): 845-54.
[16]LITTLE K M, LEE J K, LEY K. ReSASC: a resampling-based algorithm to determine differential protein expression from spectral count data [J]. Proteomics, 2010, 10(6): 1212-22.
[17]LEITCH M C, MITRA I, SADYGOV R G. Generalized Linear and Mixed Models for Label-Free Shotgun Proteomics [J]. Statistics and its interface, 2012, 5(1): 89-98.
[18]BRANSON O E, FREITAS M A. A multi-model statistical approach for proteomic spectral count quantitation [J]. Journal of proteomics, 2016, 144(23-32.
[19]XU J, WANG L, LI J. Biological network module-based model for the analysis of differential expression in shotgun proteomics [J]. Journal of proteome research, 2014, 13(12): 5743-50.
[20]ZHANG B, WANG J, WANG X, et al. Proteogenomic characterization of human colon and rectal cancer [J]. Nature, 2014, 513(7518): 382-7.
[21]PAVELKA N, FOURNIER M L, SWANSON S K, et al. Statistical similarities between transcriptomics and quantitative shotgun proteomics data [J]. Molecular & cellular proteomics : MCP, 2008, 7(4): 631-44.
[22]KIKUCHI T, HASSANEIN M, AMANN J M, et al. In-depth proteomic analysis of nonsmall cell lung cancer to discover molecular targets and candidate biomarkers [J]. Molecular & cellular proteomics : MCP, 2012, 11(10): 916-32.
[23]SHAO C, LIU Y, RUAN H, et al. Shotgun proteomics analysis of hibernating arctic ground squirrels [J]. Molecular & cellular proteomics : MCP, 2010, 9(2): 313-26.
[24]HUANG L, ZHENG D, ZALKIKAR J, et al. Zero-inflated Poisson model based likelihood ratio test for drug safety signal detection [J]. Statistical methods in medical research, 2017, 26(1): 471-88.
[25]GOSETTI F, MAZZUCCO E, ZAMPIERI D, et al. Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry [J]. Journal of chromatography A, 2010, 1217(25): 3929-37.
[26]ANNESLEY T M. Ion suppression in mass spectrometry [J]. Clinical chemistry, 2003, 49(7): 1041-4.
[27]TAYLOR P J. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry [J]. Clinical biochemistry, 2005, 38(4): 328-34.
[28]COX J, MANN M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification [J]. Nature biotechnology, 2008, 26(12): 1367-72.

[29]MUTH T, VAUDEL M, BARSNES H, et al. XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results [J]. Proteomics, 2010, 10(7): 1522-4.
[30]LYASHEVSKA O, BRUS D J, VAN DER MEER J. Mapping species abundance by a spatial zero-inflated Poisson model: a case study in the Wadden Sea, the Netherlands [J]. Ecology and evolution, 2016, 6(2): 532-43.
[31]THE UNIPROT C. UniProt: the universal protein knowledgebase [J]. Nucleic acids research, 2017, 45(D1): D158-D69.
[32]KANEHISA M, FURUMICHI M, TANABE M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs [J]. Nucleic acids research, 2017, 45(D1): D353-D61.
[33]KANEHISA M, GOTO S. KEGG: kyoto encyclopedia of genes and genomes [J]. Nucleic acids research, 2000, 28(1): 27-30.
[34]MOOTHA V K, LINDGREN C M, ERIKSSON K F, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes [J]. Nature genetics, 2003, 34(3): 267-73.
[35]ZEILEIS A, KLEIBER C, JACKMAN S. Regression models for count data in R [J]. J Stat Softw, 2008, 27(8): 1-25.
[36]ALROY I, YARDEN Y. The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions [J]. FEBS letters, 1997, 410(1): 83-6.
[37]WICKHAM H. The Split-Apply-Combine Strategy for Data Analysis [J]. J Stat Softw, 2011, 40(1): 1-29.
[38]DUNTEMAN G H, HO M-H R. An introduction to generalized linear models [M]. Thousand Oaks, Calif.: Sage Publications, 2006.
[39]WANG X, TERFVE C, ROSE J C, et al. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens [J]. Bioinformatics, 2011, 27(6): 879-80.
[40]SUBRAMANIAN A, TAMAYO P, MOOTHA V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(43): 15545-50.
[41]BENJAMINI Y, DRAI D, ELMER G, et al. Controlling the false discovery rate in behavior genetics research [J]. Behavioural brain research, 2001, 125(1-2): 279-84.
[42]HOLLAS W, BLASI F, BOYD D. Role of the urokinase receptor in facilitating extracellular matrix invasion by cultured colon cancer [J]. Cancer research, 1991, 51(14): 3690-5.
[43]STETLER-STEVENSON W G, AZNAVOORIAN S, LIOTTA L A. Tumor cell interactions with the extracellular matrix during invasion and metastasis [J]. Annual review of cell biology, 1993, 9(541-73.
[44]OWENS L V, XU L, CRAVEN R J, et al. Overexpression of the focal adhesion kinase (p125FAK) in invasive human tumors [J]. Cancer research, 1995, 55(13): 2752-5.

[45]CANCE W G, HARRIS J E, IACOCCA M V, et al. Immunohistochemical analyses of focal adhesion kinase expression in benign and malignant human breast and colon tissues: correlation with preinvasive and invasive phenotypes [J]. Clinical cancer research : an official journal of the American Association for Cancer Research, 2000, 6(6): 2417-23.
[46]BI X, LIN Q, FOO T W, et al. Proteomic analysis of colorectal cancer reveals alterations in metabolic pathways: mechanism of tumorigenesis [J]. Molecular & cellular proteomics : MCP, 2006, 5(6): 1119-30.
[47]CHEN L S, HUTTER C M, POTTER J D, et al. Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data [J]. Am J Hum Genet, 2010, 86(6): 860-71.
[48]BIENZ M, CLEVERS H. Linking colorectal cancer to Wnt signaling [J]. Cell, 2000, 103(2): 311-20.
[49]GULHATI P, BOWEN K A, LIU J, et al. mTORC1 and mTORC2 regulate EMT, motility, and metastasis of colorectal cancer via RhoA and Rac1 signaling pathways [J]. Cancer research, 2011, 71(9): 3246-56.
[50]MEES S T, MENNIGEN R, SPIEKER T, et al. Expression of tight and adherens junction proteins in ulcerative colitis associated colorectal carcinoma: upregulation of claudin-1, claudin-3, claudin-4, and beta-catenin [J]. International journal of colorectal disease, 2009, 24(4): 361-8.
[51]CORCORAN R B, EBI H, TURKE A B, et al. EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib [J]. Cancer discovery, 2012, 2(3): 227-35.
[52]LASCORZ J, FORSTI A, CHEN B, et al. Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility [J]. Carcinogenesis, 2010, 31(9): 1612-9.
[53]FANG J Y, RICHARDSON B C. The MAPK signalling pathways and colorectal cancer [J]. The Lancet Oncology, 2005, 6(5): 322-7.
[54]SUZUKI H, WATKINS D N, JAIR K W, et al. Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer [J]. Nature genetics, 2004, 36(4): 417-22.
[55]GUPTA R A, DUBOIS R N. Colorectal cancer prevention and treatment by inhibition of cyclooxygenase-2 [J]. Nature reviews Cancer, 2001, 1(1): 11-21.
[56]RODRIGUES N R, ROWAN A, SMITH M E, et al. p53 mutations in colorectal cancer [J]. Proceedings of the National Academy of Sciences of the United States of America, 1990, 87(19): 7555-9.
[57]VANAMALA J, REDDIVARI L, RADHAKRISHNAN S, et al. Resveratrol suppresses IGF-1 induced human colon cancer cell proliferation and elevates apoptosis via suppression of IGF-1R/Wnt and activation of p53 signaling pathways [J]. BMC cancer, 2010, 10(238.

[58]JOHNSON S M, GULHATI P, RAMPY B A, et al. Novel expression patterns of PI3K/Akt/mTOR signaling pathway components in colorectal cancer [J]. Journal of the American College of Surgeons, 2010, 210(5): 767-76, 76-8.
[59]DIN F V, VALANCIUTE A, HOUDE V P, et al. Aspirin inhibits mTOR signaling, activates AMP-activated protein kinase, and induces autophagy in colorectal cancer cells [J]. Gastroenterology, 2012, 142(7): 1504-15 e3.
[60]KARAPETIS C S, KHAMBATA-FORD S, JONKER D J, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer [J]. The New England journal of medicine, 2008, 359(17): 1757-65.
[61]KHAMBATA-FORD S, GARRETT C R, MEROPOL N J, et al. Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab [J]. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2007, 25(22): 3230-7.
[62]BENVENUTI S, SARTORE-BIANCHI A, DI NICOLANTONIO F, et al. Oncogenic activation of the RAS/RAF signaling pathway impairs the response of metastatic colorectal cancers to anti-epidermal growth factor receptor antibody therapies [J]. Cancer research, 2007, 67(6): 2643-8.
[63]SUGIYAMA M, TAKAHASHI H, HOSONO K, et al. Adiponectin inhibits colorectal cancer cell growth through the AMPK/mTOR pathway [J]. International journal of oncology, 2009, 34(2): 339-44.

APPENDIX

In appendix, we list some core codes for simulation study and real data sets analysis. The whole process employs parallel computation and needs the support of multi-core computer.

#Simulation

```
library(pscl)
library(MASS)
library(plyr)
library(LaplacesDemon)
library(stats)

Args <- commandArgs()
ranseeds <- Args[3]
set.seed(ranseeds)

fclist <- c(1.2, 1.5, 2.0, 3.0, 4.0)
plist <- c(0.3, 0.5, 0.8)
zero.ratio <- c(0, 0.3, 0.5, 0.7)
modulesize <- 100
modulelist <- matrix(1:modulesize, nrow=modulesize, ncol=1)

#read data
data.by <- read.table("squirrle.txt", header=TRUE, sep="\t",
quote="")
data.by <- data.by[, c(6:17)]
proteinsize <- nrow(data.by)
samplesize <- ncol(data.by)

#constaruct 2-group matrix
Group <- matrix(rep(1:0, each=proteinsize*samplesize/2),
nrow=proteinsize, ncol=samplesize)
Group <- as.factor(Group)
```

```
#construct module data, both content and length are randomly
x <- runif(modulesize, min=0.1,max=1)
x <- x * proteinsize / sum(x)
x <- round(x)
x[1] <- x[1] + proteinsize - sum(x)

sizenumber <- sample(1:proteinsize, proteinsize)
data.module <- list()
listGS <- list()
for (i in 1:modulesize) {
  data.module[[i]] <- data.frame(t(sizenumber[1:x[i]]))
  listGS[[paste("module", i, sep = "")]] <-
as.character(sizenumber[1:x[i]])
  names(listGS[[i]]) <- as.character(sizenumber[1:x[i]])
  sizenumber <- sizenumber[-(1:x[i])]
}
do.call(rbind.fill, data.module)
listGS2 <- list(listGS=listGS)
modulename <- c()
for (i in 1:modulesize) {
  modulename <- c(modulename, paste("module",i,sep=""))
}

#create output result for module list
tempmatrix <- matrix(nrow=modulesize, ncol=1)
for (i in 1:modulesize) {
  tempmatrix[i] <- paste(as.character(data.module[[i]]),
collapse=";")
}
table.module <- data.frame(ID=modulelist,
gene.symbol=tempmatrix)

#start simulation
namelist <- c(1:proteinsize)
flag <- as.numeric(data.module[[1]])
for (zr in zero.ratio) {
  #define true zero cells
  count <- data.by
  signalmatrix <- matrix(rep(0,samplesize*proteinsize),
nrow=proteinsize, ncol=samplesize)
  for (m in 1:samplesize) {
    ob.zero <- which(count[,m] == 0)
    if (round(zr*proteinsize) <= length(ob.zero)) {
```

```
      true.zero <- sample(ob.zero, round(zr*proteinsize))
    } else {
      true.zero <- sample(which(count[,m] != 0),
round(zr*proteinsize)-length(ob.zero))
      true.zero <- c(true.zero, ob.zero)
    }
    count[c(true.zero),m] = 0
    signalmatrix[c(true.zero),m] = 1
  }
  X <- as.numeric(rowSums(count))
  Y <- as.numeric(colSums(count))

  for (fc in fclist) {
    for (p in plist) {
      count1 <- count
      cat("\nconstruct analyze file under p=", p, ", fc=", fc, ",
zero ratio=", zr, ":\n")
      #construct matrix with effect by using multinomial
      flag1 <- sample(flag, round(p*length(flag)))
      for (n in 1:proteinsize) {
        true.zero <- which(signalmatrix[n,] == 1)
        proba <- Y
        proba[true.zero] = 0
        if (n %in% flag1) {
          proba[1:(samplesize/2)] <- proba[1:(samplesize/2)] * fc
        }
        count1[n,] = t(rmultinom(1,sum(count[n,]),proba))
      }

      #wilcoxon rank-sum test followed by GSEA
      cat("----------wilcoxon----------\n")
      pvalue.wilcoxon <- c()
      for (i in 1:nrow(count1)) {
        pvalue.wilcoxon <- c(pvalue.wilcoxon,
wilcox.test(as.numeric(count1[i,c(1:6)]),
as.numeric(count1[i,c(7:12)]))$p.value)
      }
      pvalue.wilcoxon <- p.adjust(pvalue.wilcoxon, method = "fdr")
      names(pvalue.wilcoxon) <- as.character(1:proteinsize)
      hits <- names(pvalue.wilcoxon)[which(pvalue.wilcoxon <
0.05)]
      if (length(hits) == 0) {
        pvalue.GSEA <- matrix(rep(NA,modulesize), ncol=1)
```

```
        pvalue.GSEA.fdr <- matrix(rep(NA,modulesize), ncol=1)
      } else {
        gsca <- new("GSCA", listOfGeneSetCollections=listGS2,
geneList=pvalue.wilcoxon, hits=hits)
        gsca <- preprocess(gsca, initialIDs="Entrez.gene",
keepMultipleMappings=TRUE,
duplicateRemoverMethod="max",orderAbsValue=FALSE)
        gsca <- analyze(gsca, para=list(pValueCutoff=0.05,
pAdjustMethod="fdr", nPermutations=10, minGeneSetSize=3,
exponent=1))
        pvalue.GSEA <-
as.numeric(gsca@result$GSEA.results$listGS[,3][modulename])
        pvalue.GSEA.fdr <-
as.numeric(gsca@result$GSEA.results$listGS[,3][modulename])
      }

      #construct module matrix, then calculate p-value using model
      cat("----------model----------\n")
      count1 <- as.numeric(as.matrix(count1))
      modulematrix <- matrix(rep(0, proteinsize),
nrow=proteinsize, ncol=1)
      pvalue.NB <- matrix(nrow=modulesize,ncol=1)
      pvalue.ZIP <- matrix(nrow=modulesize, ncol=1)
      pvalue.ZINB <- matrix(nrow=modulesize, ncol=1)
      reportPoint <- c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
      for (i in 1:modulesize) {
        if ((i*100/modulesize) >= reportPoint[1]) {
          cat(" --", reportPoint[1], "%")
          reportPoint <- reportPoint[-1]
        }
        genes <- unlist(data.module[[i]])

        for (j in 1:proteinsize) {
          if (length(genes) == 0)
            break
          for (k in 1:length(genes)) {
            if (j == genes[k]) {
              modulematrix[j] <- 1
              genes <- genes[-k]
              break
            }
          }
        }
```

```
        Module <- modulematrix[,rep(1,samplesize)]
        Module <- as.factor(Module)

        fm1 <- glm.nb(count1 ~ Group*Module)
        pvalue.NB[i] <- summary(fm1)$coefficients[4,4]

        fm2 <- try(zeroinfl(count1 ~ Group*Module, dist="negbin"),
silent=TRUE)
        if ('try-error' %in% class(fm2)) {

        } else pvalue.ZINB[i] <-
summary(fm2)[1]$coefficients$count[4,4]

        fm4 <- try(zeroinfl(count1 ~ Group*Module, dist="pois"),
silent=TRUE)
        if ('try-error' %in% class(fm4)) {

        } else pvalue.ZIP[i] <-
summary(fm4)[1]$coefficients$count[4,4]

        modulematrix <- matrix(rep(0, proteinsize),
nrow=proteinsize, ncol=1)
      }

      pvalue.NB.fdr <- p.adjust(pvalue.NB, method='fdr')
      pvalue.ZIP.fdr <- p.adjust(pvalue.ZIP, method='fdr')
      pvalue.ZINB.fdr <- p.adjust(pvalue.ZINB, method='fdr')

      output.data <- data.frame(modulelist, pvalue.NB,
pvalue.NB.fdr, pvalue.ZIP, pvalue.ZIP.fdr, pvalue.ZINB,
pvalue.ZINB.fdr, pvalue.GSEA, pvalue.GSEA.fdr, table.module[,2])
      colnames(output.data) <- c("Module", "NB", "NB.fdr", "ZIP",
"ZIP.fdr", "ZINB", "ZINB.fdr", "GSEA", "GSEA.fdr", "Gene.Symbol")
      write.table(output.data, paste("/squirrel_pathway_", zr*10,
"_", p*10, fc*10, "_for_", ranseeds, "_times.txt", sep=""),
quote=FALSE, sep="\t", row.names=FALSE, col.names=TRUE,
append=FALSE)
    }
  }
}
```

#Real Data Sets Analysis

```
library(pscl)
library(MASS)
library(plyr)
library(parallel)
library(HTSanalyzeR)
library(GSEABase)
library(cellHTS2)

data.raw <-
read.table("protein_7244_normal30_tumor90_quantileLog_sorted.t
xt", header=TRUE, sep="\t", quote="", row.names=1)
data.cc <- data.raw[,c(31:120,1:30)]
count <- as.numeric(round(as.matrix(2 ^ data.cc - 1)))
Group <- as.factor(matrix(c(rep(1,90),rep(0,30)),
nrow=1)[rep(1,nrow(data.cc)),])
data.pathway <- read.table("pathway.txt", header=FALSE, sep="\t",
quote="")

#==============================model
analyze=======================================
calcu <- function(i) {
  genelist <- as.character(data.pathway[i,2])
  genes <- unlist(strsplit(genelist, ";"))
  totalnum <- length(genes)
  matchnum <- 0
  pathwaymatrix <- matrix(rep(0, nrow(data.cc)),
nrow=nrow(data.cc), ncol=1)

  for (j in 1:nrow(data.cc)) {
    if (length(genes) == 0)
      break
    gene <- row.names(data.cc[j,])
    for (k in 1:length(genes)) {
      if (gene == genes[k]) {
        pathwaymatrix[j] <- 1
        matchnum <- matchnum + 1
        genes <- genes[-k]
        break
      }
```

```
    }
  }

  if (sum(pathwaymatrix == 1) == 0) {
    NB <- NA
    ZIP <- NA
  } else {
  #calculate p-value
    Pathway <- as.factor(pathwaymatrix[,rep(1,ncol(data.cc))])
    fm1 <- glm.nb(count ~ Group*Pathway)
    NB <- summary(fm1)$coefficients[4,4]

    fm2 <- try(zeroinfl(count ~ Group*Pathway, dist="pois"),
silent=TRUE)
    if ('try-error' %in% class(fm2)) {
      ZIP <- NA
    } else ZIP <- summary(fm2)[1]$coefficients$count[4,4]
  }

  return(c(as.character(data.pathway[i,1]), NB, ZIP, matchnum,
totalnum))
}

#do parallel calculation
cat("============model============\n")
cl <- makeCluster(getOption("cl.cores", 20))
clusterEvalQ(cl, c("pscl", "MASS"))
clusterExport(cl, c("data.pathway", "data.cc", "count", "Group",
"glm.nb", "zeroinfl"))
results <- parLapply(cl, 1:nrow(data.pathway), calcu)
res.df <- do.call('rbind',results)
NB.fdr <- p.adjust(res.df[,2], method="fdr")
ZIP.fdr <- p.adjust(res.df[,3], method="fdr")
stopCluster(cl)

#==============================wilcoxon+GSEA================
===
cat("============GSEA============\n")
listGS <- list()
for (i in 1:nrow(data.pathway)) {
  listGS[[as.character(data.pathway[i,1])]] <-
as.character(unlist(strsplit(as.character(data.pathway[i,2]),
";")))
```

```
  names(listGS[[i]]) <-
as.character(unlist(strsplit(as.character(data.pathway[i,2]),
";")))
}
listGS2 <- list(listGS=listGS)

pvalue <- matrix(rep(0,nrow(data.cc)),ncol=1)
for (i in 1:nrow(data.cc)) {
  pvalue[i] <- wilcox.test(as.numeric(data.cc[i,c(1:90)]),
as.numeric(data.cc[i,c(91:120)]))$p.value
}
p.fdr <- p.adjust(pvalue,method="fdr")
names(p.fdr) <- as.character(row.names(data.raw))
hits <- names(p.fdr)[which(p.fdr < 0.05)]
gsca <- new("GSCA", listOfGeneSetCollections=listGS2,
geneList=p.fdr, hits=hits)
gsca <- preprocess(gsca, initialIDs="Entrez.gene",
keepMultipleMappings=TRUE,
duplicateRemoverMethod="max",orderAbsValue=FALSE)
gsca <- analyze(gsca, para=list(pValueCutoff=0.05,
pAdjustMethod="fdr", nPermutations=10, minGeneSetSize=3,
exponent=1))
pvalue.GSEA <-
as.numeric(gsca@result$GSEA.results$listGS[,2][res.df[,1]])
pvalue.GSEA.fdr <-
as.numeric(gsca@result$GSEA.results$listGS[,3][res.df[,1]])


output <- data.frame(res.df[,1], res.df[,2], NB.fdr, res.df[,3],
ZIP.fdr, pvalue.GSEA, pvalue.GSEA.fdr, res.df[,4], res.df[,5])
colnames(output) <- c("Pathway", "NB", "NB.fdr", "ZIP", "ZIP.fdr",
"GSEA", "GSEA.fdr", "match", "total amount")
write.table(output, "output_CRC.txt", quote=FALSE, sep="\t",
row.names=FALSE)
```

ACKNOWLEDGEMENT

Thanks to Shanghai Jiao Tong University for offering me a platform to improve myself. During the research I learned so many knowledge and equipped myself with all kinds of skills.

Thanks to Prof. Jing Li and all other teachers in the department of Bioinformatics & Biostatistics. During the whole thesis process, their supervision and guidance is what keeps me progressing. And owing to their kindly guide, I can overcome so many problems to finish this paper.

Thanks to Bo Wang, Jie Ren and Xi Cheng from Jing Li's Lab. I could not handle the task without their assistance.