Computational Biology and Chemistry 67 (2017) 255–259

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

A Gibbs sampling method to determine biomarkers for asthma

Zhi-Jian Huang[a], Qin-Hai Shen[b], Yan-Sheng Wu[c], Ya-Li Huang[d,*]

[a] Department of Emergency, Xiamen Hospital of Traditional Chinese Medicine, Xiamen 361009, Fujian, PR China
[b] Department of Medicine, Shandong Medical College, Jinan, 250002, Shandong, PR China
[c] Department of Spine (Second), Traditional Chinese Medical Hospital of Xinjiang Uygur Autonomous Region, Urumqi, 830000, Xinjiang, PR China
[d] Nuclear Medicine Department, Qilu Hospital of Shandong University, NO. 107 Wenhua West Road, Jinan, 250012, Shandong PR China

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form 22 December 2016
Accepted 18 January 2017
Available online 22 January 2017

Keywords:
Asthma
Gibbs sampling
Molecular function
Markov chain
Pathway enrichment analysis

ABSTRACT

Purpose: To identify potential biomarkers and to uncover the mechanisms underlying asthma based on Gibbs sampling.
Methods: The molecular functions (MFs) with genes greater than 5 were determined using AnnotationMFGO of BAGS package, and the obtained MFs were then transformed to Markov chain (MC). Gibbs sampling was conducted to obtain a new MC. Meanwhile, the average probabilities of MFs were computed via MC Monte Carlo (MCMC) algorithm, followed by identification of differentially expressed MFs based on the probabilities of MF more than 0.6. Moreover, the differentially expressed genes (DEGs) and their correlated genes were screened and merged, called as co-expressed genes. Pathways enrichment analysis was implemented for the co-expressed genes.
Results: Based on the gene set more than 5, overall 396 MFs were determined. After Gibbs sampling, 5 differentially expressed MF were acquired according to alfa.pi > 0.6. Moreover, the genes in these 5 differentially expressed MF were merged, and 110 DEGs were identified. Subsequently, 338 co-expressed genes were gained. Based on the P value < 0.01, the co-expressed genes were significantly enriched in 6 pathways. Among these, ubiquitin mediated proteolysis contained the maximum numbers of 35 co-expressed genes, and cell cycle were enriched by the second largest number of 11 co-expressed genes, respectively.
Conclusions: The identified pathways such as ubiquitin mediated proteolysis and cell cycle might play important roles in the development of asthma and may be useful for developing the credible therapeutic approaches for diagnosis and treatment of asthma in future.

1. Introduction

Asthma, characterized by bronchospasm, airflow obstruction and hyperresponsiveness, is a complex chronic inflammatory disorder of the airways (Maddox and Schwartz, 2002). Its symptoms include wheezing, shortness of breath, coughing, and chest tightness. Of note, these symptoms are often worse in the early morning, at night, or in response to exercise (Network, 2008). Worriedly, the asthma prevalence is increasing significantly in the world. As in documented, the occurrence rate of asthma in children in Britain was 10.2% in 2000 while 20.9% in 2011 (Koehoorn et al., 2013). Moreover, asthma exacerbations are the most common causes of hospitalization and places a heavy burden on society (Matterne et al., 2011). However, the pathogenesis is still poorly understood. Therefore, systemic identification of biomarkers is desirable for the rational management of asthma.

Recently, advances have increased the understanding of the complex mechanisms underlying asthma. As reported, asthma has been demonstrated to be caused by a combination of genetic as well as environmental interactions with low-level and ubiquitous environmental exposures (Gibson, 2008; Miller and Ho, 2008). For example, the signal transduction pathway of Janus kinase (JAK)-signal transducer and activator of transcription (STAT) is suggested to be involved in the asthma development (Pfitzner et al., 2004). Moreover, Li and colleagues have indicated that NF-κB signaling pathway is participated in the asthmatic change driven by CKLF1 (Li et al., 2014). IL13 mRNA is increased in the bronchial tissue of nonatopic and atopic asthmatic patients (Humbert et al., 1997). Another study has demonstrated that VEGF is increased in patients with asthma and is associated with the disease severity (Voelkel et al., 2006). Nevertheless, it is badly needed for researchers to better understand the pathological mechanism of asthma and customized therapies, detecting molecular functions (MFs) on the

* Corresponding author.
E-mail address: huangyaliHE@163.com (Y.-L. Huang).

http://dx.doi.org/10.1016/j.compbiolchem.2017.01.008
1476-9271/

basis of significant genes may be helpful to solve this difficulty to a certain degree. Gene ontology (GO) analysis, in a flexible and dynamic manner, has been widely employed as functional enrichment studies for large-scale genes (Ashburner et al., 2000). As we all know, GO includes 3 categories: biological process (BP), cellular component (CC), and MF. MF is described as the biological activity of genes, and MF also shows what is done without specifying when or where the events happen (Ashburner et al., 2000). Thus, if one aimed to study the biological activities of a given gene at the molecular level, the best method was to analyze the corresponding MFs of this gene. Unfortunately, few studies focused on the functions of significant MFs in the disease progression.

Gibbs sampling is frequently applied to be a means of statistical inference, particularly, Bayesian inference (Walsh, 2004). Significantly, Gibbs sampling, is a Markov Chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution (Chib and Winkelmann, 2012; El-Hay et al., 2012; Kozumi and Kobayashi, 2011). Significantly, based on the probabilities, differentially expressed function terms and key genes are likely to be identified which might be important to reveal the disorder pathology. Thus, in our work, we applied Gibbs sampling to investigate significances of MFs and their roles in asthma.

Hence, in the current study, we selected the differentially expressed MFs in asthma using Gibbs sampling based on Bayesian approaches and MCMC. The MFs with genes greater than 5 were determined using AnnotationMFGO of BAGS package, and the obtained MFs were then transformed to Markov Chain (MC). Next, Gibbs sampling was carried out to obtain a new MC. At the same time, the average probability of MFs were computed by means of MCMC algorithm, followed by the identification of differentially expressed MFs based on the probabilities of MF more than 0.6. Moreover, the differentially expressed genes (DEGs) and their correlated genes were screened and merged, called as co-expressed genes. Pathways enrichment analysis was implemented for the co-expressed genes. Our findings will help us to understand the molecular mechanisms underlying asthma.

2. Materials and methods

2.1. Microarray data and data processing

The profile E-GEOD-35571 (Williams-DeVane et al., 2013) was downloaded from EMBL-EBI which is a public-functional database. The platform was Affymetrix Human Genome U133 Plus 2.0 Array. In the gene microarray data of E-GEOD-35571, there were 131 peripheral blood samples from patients, including 60 patients with asthma, 64 samples without asthma as controls, and 7 samples which were not available. In order to reveal the molecular mechanisms of asthma, we only selected 60 patients with asthma, 64 samples without asthma for further analysis. The raw data and the probe annotation files were downloaded for further analysis.

The probe data were processed to corresponding gene symbols on the basis of the annotation of platform. Next, the average expression value was computed if there were several probes mapping to one gene symbol. Then, the gene expression matrix was collected.

2.2. Gibbs sampling

2.2.1. General sampling strategy

Before performing the Gibbs sampling, the introduction of MC was in order. In such a sampling, Y_x means the value of a random variable at time x, and state space denotes the range of all possible Y values. This random variable is a Markov process when the transition probabilities between different values in the state space only rely on the present state of random variable, that is

$$\text{probability}\ (Y_{x+1} = s_B \mid Y_0 = s_K, \cdots, Y_x = s_A) = \text{probability}\ (Y_{x+1} = s_B \mid Y_x = s_A)$$

Hence, for a Markov random variable, the information with respect of the past state needed to anticipate the future is the present state of the random variable, knowledge of the values of earlier states do not alter the transition probability. A MC refers to a sequence of random variables $(Y_0, \cdots, Y_n)$ produced by a Markov process and is defined most critically by the transition probabilities, probability (A, B) = probability ($A \rightarrow B$), which stands for the probability that a process at state space s_A moves to other state s_B. The specifically formula was as follows:

$$\text{probability}\ (A, B) = \text{probability}\ (A \rightarrow B) = \text{probability}\ (Y_{x+1} = s_B \mid Y_x = s_A).$$

In statistics, based on Bayesian analysis, MCMC approach governed by the desired posterior probability is one category of algorithms for sampling from a probability distribution according to establishing a MC which has the desired distribution as the equilibrium distribution. Then, the state of the chain is used as a sample of the desired distribution. The formula of posterior probability was shown:

$$\text{probability}\ (A \cap B) = \text{probability}\ (A)^* \text{probability}\ (B|A) = \text{probability}\ (B)^* \text{probability}\ (A|B).$$

Where probability (A) is the prior probability of A, probability (B) is the prior probability of B or normalized constant, probability (B|A) is the posterior probability of B, probability (A|B) is the posterior probability of A.

Posterior probability = (likelihood * prior probability)/normalized constant,

Moreover, probability (B|A)/probability (B) is standardised likelihood.

Thus, posterior probability = standardised likelihood * prior probability.

2.2.2. Identification of MC dataset

To performed the Gibbs sampling, the microarray dataset were transformed into MC dataset. First of all, we calculated the average expression of each gene of samples between the two groups, which was defined as A1 and A2. Then, the mean value of A1 and A2 was also computed, called as overall average. By comparing the average expression of each gene of samples between the two groups and overall average, the genes showing changes in the average expression value and general average were found. Next, AnnotationMFGO of a Bayesian approach for geneset selection (BAGS) package (Quiroz-Zarate et al., 2013) was utilized to map these obtained genes to the MFs. MFs with genes greater than 5 were determined. Subsequently, MF dataset were converted to a MC dataset using MCMCDataSet function of BAGS package combined with sample information.

2.2.3. Computing the probabilities of MF

The pivotal issue to the Gibbs sampling is that one only considers univariate conditional probability distributions. In our study, we firstly defined an empty set. Then, the MC dataset obtained above including MFs with genes greater than 5 (count = N) were deposited to this empty set. Subsequently, Gibbs sampling was implemented to construct the 10000 dimensional random vectors of N samples. Next, these 10000 dimensional random vectors were initialized. Among these, one vector was extracted each time to generate the random number. Repeating this process 10000 times, a new MC, i.e. 10000 probability of each MF, was