Computational Biology and Chemistry 70 (2017) 15–20

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Screening disrupted molecular functions and pathways associated with clear cell renal cell carcinoma using Gibbs sampling

Ning Nan, Qi Chen, Yu Wang, Xu Zhai, Chuan-Ce Yang, Bin Cao, Tie Chong*

Department of Urinary Surgery, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710004, China

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form 28 June 2017
Accepted 9 July 2017
Available online 13 July 2017

Keywords:
Clear cell renal cell carcinoma
Molecular functions
Gibbs sampling
Markov chain
Oxidoreductase activity
Oxidative phosphorylation

ABSTRACT

Objective: To explore the disturbed molecular functions and pathways in clear cell renal cell carcinoma (ccRCC) using Gibbs sampling.
Methods: Gene expression data of ccRCC samples and adjacent non-tumor renal tissues were recruited from public available database. Then, molecular functions of expression changed genes in ccRCC were classed to Gene Ontology (GO) project, and these molecular functions were converted into Markov chains. Markov chain Monte Carlo (MCMC) algorithm was implemented to perform posterior inference and identify probability distributions of molecular functions in Gibbs sampling. Differentially expressed molecular functions were selected under posterior value more than 0.95, and genes with the appeared times in differentially expressed molecular functions $\geq$5 were defined as pivotal genes. Functional analysis was employed to explore the pathways of pivotal genes and their strongly co-regulated genes.
Results: In this work, we obtained 396 molecular functions, and 13 of them were differentially expressed. Oxidoreductase activity showed the highest posterior value. Gene composition analysis identified 79 pivotal genes, and survival analysis indicated that these pivotal genes could be used as a strong independent predictor of poor prognosis in patients with ccRCC. Pathway analysis identified one pivotal pathway – oxidative phosphorylation.
Conclusions: We identified the differentially expressed molecular functions and pivotal pathway in ccRCC using Gibbs sampling. The results could be considered as potential signatures for early detection and therapy of ccRCC.

1. Introduction

Renal cell cancer (RCC) is the 16th most leading cause of death from malignancy globally (Znaor et al., 2015). Clear cell RCC (ccRCC) is the most common histologic subtype, representing 75–85% of all RCC (Ljungberg et al., 2010). What is worse, its 5-year survival rate is only 50–69% (Gudbjartsson et al., 2005). Over the past decade, despite advances on some new agents, most patients with advanced ccRCC progress on therapy, but the prognosis is still poor (Srinivasan et al., 2015).

Many genetic alterations have been observed to be association with ccRCC. Miller et al. (Miller et al., 2015) found that reduced levels of *LARGE2* and *ISPD* associated with increased mortality in patients with ccRCC. Li et al. (Li et al., 2015a) indicated that high *KDM6 B* level was positively involved in poor ccRCC prognosis, and the knockdown of *KDM6 B* could inhibit ccRCC tumorigenesis *in vitro*. Wan et al. (Wan et al., 2015) suggested that up-regulation expression of *COL6A1* correlated with poor prognosis in ccRCC patients and stimulated tumor growth *in vivo*. However, most of the biological processes in cells arises from complex synergistic actions among genes, the genes alone may lead to false positives (Glazko and Emmert-Streib, 2009). Disclosing molecular functions based on gene ontology analysis may contribute to understanding how gene perturbations account for the pathogenic procedure of complex diseases. With advances on high-throughput experimental technologies, researches have focused on post-genomic analysis to reveal disease-related genes and pathways associated with the underlying mechanism of complex diseases (Yang et al., 2014; Feng et al., 2015).

Gibbs sampling is a commonly used randomized algorithm for statistical inference, especially Bayesian inference. It is a calculation method of Markov Chain Monte Carlo (MCMC) algorithm for calculating numerical approximations of multi-dimensional integrals (Walsh, 2004). MCMC method can be used to obtain a sequence of observations that is a valid approximation of the

* Corresponding author at: Department of Urinary Surgery, The Second Affiliated Hospital of Xi'an Jiaotong University, No. 157, Xiwu Road, Xi'an, 710004, China.
E-mail address: chongtie86@126.com (T. Chong).

http://dx.doi.org/10.1016/j.compbiolchem.2017.07.002
1476-9271/

posterior probability distributions (Huelsenbeck and Ronquist, 2001). Bayesian inference shares the key feature of combining prior belief with the available data to produce statistically consistent samples from a posterior distribution, rather than searching for a single well-scoring model (Voelz and Zhou, 2014; Habeck et al., 2005). Based on the probability distributions, the researchers could obtain pivotal genes or pathways that might be associated with the disorder pathology of complex diseases.

In this paper, we attempted to identify disturbed molecular functions and pathways in ccRCC using Gibbs sampling based on Bayesian inference and MCMC algorithm. The performance of Gibbs sampling in this simulation framework can guide the choice of methods for disturbed molecular identification using genetics data.

2. Materials and methods

2.1. Data collection and processing

The gene expression profile, under the accession number of E-GEOD-40435 (Wozniak et al., 2013), was downloaded from ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/), using Illumina HumanHT-12 v4 Expression BeadChips. A total of 101 pairs of ccRCC tumors and adjacent non-tumor renal tissue biopsies were collected from patients, which included 42 females and 59 males. There were no significant differences between male and female individuals for known ccRCC risk factors. The detailed clinical and demographic characteristics of all samples have been described in the original study (Wozniak et al., 2013).

In order to eliminate the influence of nonspecific hybridization, background correction and normalization were carried out via robust multichip average method (Ma et al., 2006) and quantile based algorithm (Rifai and Ridker, 2001), respectively. Perfect match and mismatch value were revised by Micro Array Suite 5.0 algorithm (Pepper et al., 2007). The probe annotation data were downloaded for further analysis. Each probe is mapped to one gene, and the probe is discarded if it can't match any one. There were 47,231 probes presented in the dataset. After data preprocessing, we obtained 31,314 genes ultimately.

2.2. Gibbs sampling

2.2.1. Identifying molecular functions

Gene Ontology (GO) is a collaborative bioinformatics resource that describes gene product functions using ontologies to represent biological knowledge (Gene Ontology Consortium, 2015). Go project has developed three independent ontologies: biological processes, cellular components and molecular functions (Ashburner et al., 2000). Molecular function describes the molecular activities of gene products, including specific binding to ligands or structures. It only defines what is done without expounding where or when the event actually occurs (Ashburner et al., 2000). To identify disrupted molecular functions using Gibbs sampling, the dataset needs to be transformed to a data set with functional class expression measurements (Quiroz-Zárate et al., 2013). Here, the average expression of each gene was calculated in two conditions (ccRCC vs controls) respectively, and the overall mean value of each gene across two conditions was also computed. By comparing the average expression of each gene in specific condition with overall mean value, the genes showing expression changes were obtained. Then, gene molecular functions of expression changed genes were classed to GO project using AnnotationMFGO function of a Bayesian approach for geneset selection (BAGS) package. The BAGS package provides functions to perform statistical identification of gene functional classes that behave in a distinct manner between the phenotypes of interest for datasets under cross-sectional or time series designs (Quiroz-Zárate et al., 2013). In this analysis, only molecular functions with at least 5 genes were retained for further analysis. Here we obtained the data set containing 396 molecular functions.

2.2.2. Markov chains

Gibbs sampling is MCMC algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution. In order to perform Gibbs sampling, the above molecular functions should be converted into Markov chains, which is a sequence of random variables where the distribution of each random variable depends only on the value of the previous random variables (Hastings, 1970). Here, combining the sample information, all molecular functions were converted into Markov chains via MCMC DataSet function in BAGS package.

2.2.3. Posterior inference

After transforming the above molecular functions into Markov chains, its posterior inference was used to identify probability distributions of molecular functions from ccRCC (Moradkhani et al., 2012). In Bayesian statistics, the recent development of MCMC method has been a key step in making it possible to compute large hierarchical models that require integrations over hundreds or even thousands of unknown parameters (Banerjee et al., 2014). Given the observations, $m=(m_1, \ldots, m_n)$, the posterior distribution of β, $\pi(\beta|m)$ is given by:

$$\pi(\beta|m) \propto L(m|\beta)p(\beta)$$

Where $p(\beta)$ is the prior distribution of β and $L(m|\beta)$ is the likelihood function written as:

$$L(m|\beta) = p^n(1-p)^n exp\{-\sum_i \rho_p(m_i - x_i'\beta)\}$$

2.3. Probabilities of molecular functions

Gibbs sampling generates a Markov chains of samples, each of which is correlated with nearby samples. Owing to significantly complicated samples in this task, we adopted the Gibbs sampling to observe the multivariate distribution in complex samples. It is not only to fit the complicated samples, but additionally to use the fit to guide future predictions. In the present research, in order to implement the Gibbs sampler, we need to define an empty object in which the above Markov chains data set was to be kept. Moreover, we performed the Gibbs sampling through the Gibbs2 function of BAGS package, which provided the Markov chains for the parameters of interest that would form their posterior distribution, and this function could be used to obtain the gene sets that were differentially expressed between control and disease groups (Quiroz-Zarate and Quiroz-Zarate, 2013). Gibbs sampling produces samples of functions of interest whose distribution converges to posterior distribution. Specific steps were as follows: Firstly, an empty Gibbs sampling set was defined, and the above identified molecular functions (396 samples) was deposited to this empty Gibbs sampling set. Next, Gibbs sampling was employed to construct k-dimensional ($k=10{,}000$) random vectors of 396 samples. Meanwhile we initialized the k-dimensional vectors randomly and selected the remaining one element by fixing k-1 elements of the vectors. Repeating k times, we generated a new Markov chain. Here, there were a total of 10,000 iterations for the Markov chains, and iterations 2000–10,000 showed satisfactory convergence of all parameters in all chains. The probability of an molecular function appearing k times, i.e. the posterior value, was calculated according to the formula: