Genomics Proteomics Bioinformatics 11 (2013) 345–353

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Identification of Immunity-related Genes in *Arabidopsis* and Cassava Using Genomic Data

Luis Guillermo Leal [1,#], Álvaro Perez [2,#], Andrés Quintero [2], Ángela Bayona [2], Juan Felipe Ortiz [2], Anju Gangadharan [3], David Mackey [3], Camilo López [2], Liliana López-Kleine [1,*]

[1] *Department of Statistics, Universidad Nacional de Colombia, Bogotá 111321, Colombia*
[2] *Department of Biology, Universidad Nacional de Colombia, Bogotá 111321, Colombia*
[3] *Department of Molecular Genetics, The Ohio State University, Columbus, OH 43210, USA*

Received 12 June 2013; revised 19 September 2013; accepted 22 September 2013
Available online 6 December 2013

KEYWORDS

Arabidopsis;
Cassava;
Functional gene prediction;
Genomic data;
Kernel canonical correlation analysis;
Plant immunity

Abstract Recent advances in genomic and post-genomic technologies have provided the opportunity to generate a previously unimaginable amount of information. However, biological knowledge is still needed to improve the understanding of complex mechanisms such as plant immune responses. Better knowledge of this process could improve crop production and management. Here, we used holistic analysis to combine our own microarray and RNA-seq data with public genomic data from *Arabidopsis* and cassava in order to acquire biological knowledge about the relationships between proteins encoded by immunity-related genes (IRGs) and other genes. This approach was based on a kernel method adapted for the construction of gene networks. The obtained results allowed us to propose a list of new IRGs. A putative function in the immunity pathway was predicted for the new IRGs. The analysis of networks revealed that our predicted IRGs are either well documented or recognized in previous co-expression studies. In addition to robust relationships between IRGs, there is evidence suggesting that other cellular processes may be also strongly related to immunity.

Introduction

Recent advances in genomic and post-genomic technologies have provided the opportunity to generate vast datasets. However, the data stored in genomic databases does not itself provide an understanding of biological processes and has not always been generated under biological conditions of interest. Nevertheless, available data could be combined with own data generated in-house for the biological condition of interest to improve results and generate more confident biological conclusions. The new challenge is to develop mathematical methods

* Corresponding author.
E-mail: llopezk@unal.edu.co (López-Kleine L).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.09.010

to assess biological problems or phenomena from a holistic or system-level perspective and to use own and other information available. Approaches to extract knowledge from genomic databases and combine these data with new experimental data should allow the integration, interpretation and analysis of genomic and post-genomic data and should represent the acquired biological knowledge in the form of gene or protein networks showing functional/co-expression relationships or other structured representation. This representation should reflect relationships at the individual and categorical levels, which would assemble genes/proteins of known, unknown and hypothetical functions.

Several different approaches have been developed in recent years to assess relationships between functionally known and unknown genes/proteins through biological networks and predict new functions of genes/proteins, especially in humans [1]. These methods are often supervised and allow the integration of multiple genomic data sources in different ways [2,3], thus generating reliable and robust results, often including the prediction of new protein functions [4]. Due to the specific and complicated characteristics of genomic data, proper analysis and generation of useful inference represent real mathematical and statistical challenges.

Predictions of function are better conducted using methods that allow the integration of prior knowledge (supervised methods), the identification of non-linear relationships and the fusion of heterogeneous genomic and post-genomic data. Kernel methods [5] have these characteristics and among them, kernel canonical correlation analysis (KCCA) can be useful in relating proteins of known function with those of unknown function to predict participation in processes of interest. Earlier studies have reported the use of KCCA methods to predict the functions of unknown proteins [4,6,7]. KCCA offers a rigorous mathematical but also intuitive framework to represent biological data through kernel functions [4,5]. KCCA provides a methodology for supervised network inference and does not require exhaustive data assumptions [8]. It is therefore in contrast to alternative strategies such as Naïve Bayes (NB) models [9], which require regularization methods and have challenges of computational efficiency in the presence of many data sets [10].

Losses caused by plant pathogens represent one of the most important limitations in crop production, which can compromise the food supply [11]. Plant immunity depends on the recognition of conserved microbe-associated molecular patterns (MAMPs) or strain-specific effectors by pattern recognition receptors (PPRs) or resistance (R) proteins, leading to MAMP-triggered immunity (MTI) and effector-triggered immunity (ETI), respectively [11,12]. Upon recognition, plants activate a complex network of responses that includes signal transduction pathways, novel protein interactions and coordinated changes in gene expression [13]. Detailed information concerning specific and punctual interactions between effector and resistance proteins has been accumulated in the recent years; in some cases, a global picture for some of these interactions has been established [9,14]. Immunity networks have been described for model plants such as *Arabidopsis* and rice primarily using yeast-two hybrid experiments [15,16].

In this study, we employed a kernel-based approach to reconstruct functional relationships between genes based on genomic and post-genomic data from various sources (primarily extracted from databases but also produced by laboratory experiments) for a group of well-characterized immunity-related genes (IRGs). We employed this approach to analyze *Arabidopsis* and cassava (*Manihotesculenta*), a staple crop with little genomic information available, following challenge with bacterial pathogens. This approach allowed us to identify a group of new IRGs in both species. Many of the identified genes were of unknown function. Based on our further detailed analyses and literature knowledge, we established a list of top gene candidates potentially related to immune responses. These results indicate that publically-available data can be combined with in-house generated data using novel data-mining methods to potentially answer challenging biological questions.

Results

Exploratory analysis of categorical data

A total of 22 datasets were collected for *Arabidopsis* and cassava (see Materials and methods section for more details). Number of genes and the number of columns for each dataset are listed in Tables S1 and S2. To obtain a preliminary architecture of the data, we conducted classical descriptive multivariate analyses using multiple correspondence analysis (MCA), clustering and principal component analysis (PCA) [17] as a first step to evaluate the data structure, reveal unknown relationships and reveal clusters of genes potentially involved in immune responses. Our results showed that no groups of IRGs were clearly detected, indicating that functional relationships cannot be extracted using linear descriptive methods. Nevertheless, we were able to summarize the information of microarray data with fewer variables using an exploratory descriptive analysis. We found that most of the information contained in the microarrays is correlated and can be represented with two new variables (principal components). Accordingly, only a small portion of genes have different expression behaviors across experiments, which could be new IRGs. Furthermore, we found that RNA-seq data contains information that complements the microarray data. These results are useful and indicate that expression data contains valuable information to differentiate IRGs from non-IRGs if a more appropriate method is implemented.

All in all, the exploratory analyses showed that IRGs cannot be grouped together using only linear methods and methods such as KCCA (introduced in following section) are desired. For details on the procedure and the results of exploratory analyses, see the Supplementary File 1.

Relationship between genes/proteins obtained using KCCA

Since linear relationships between gene expression variables did not show any structure or pattern that allowed the grouping of IRGs based on either categorical or continuous data, we used non-linear kernel methods to integrate both types of data for extraction of relationships between genes. We used the supervised KCCA method [6] to predict functional relationships between genes. To do this, two reference datasets were used in the KCCA, including the real reference dataset and a random reference dataset of IRGs constructed by randomly placing a similar number of IRGs from the real reference in five categories to emulate five types of IRGs.