BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 18 2010, pages 2349–2351
doi:10.1093/bioinformatics/btq427

Systems biology

Advance Access publication July 21, 2010

A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets

André Fujita[1,*], Kaname Kojima[2], Alexandre G. Patriota[3], João R. Sato[4], Patricia Severino[5] and Satoru Miyano[1,2]

[1]Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, [2]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [3]Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, São Paulo 05508-090, [4]Center of Mathematics, Computation and Cognition, Universidade Federal do ABC, Rua Santa Adélia 166, Santo André 09210-170 and [5]Center for Experimental Research, Albert Einstein Research and Education Institute, Av. Albert Einstein, 627 - São Paulo, São Paulo 05652-000, Brazil

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We propose a likelihood ratio test (LRT) with Bartlett correction in order to identify Granger causality between sets of time series gene expression data. The performance of the proposed test is compared to a previously published bootstrap-based approach. LRT is shown to be significantly faster and statistically powerful even within non-Normal distributions. An R package named gGranger containing an implementation for both Granger causality identification tests is also provided.

Availability: http://dnagarden.ims.u-tokyo.ac.jp/afujita/en/doku.php?id=ggranger.

Contact: andrefujita@riken.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 5, 2010; revised on July 4, 2010; accepted on July 18, 2010

1 INTRODUCTION

Temporal association between genes in time series expression data can be studied through the identification of Granger causality (Fujita *et al*., 2009; Guo *et al*., 2008; Mukhopadhyay and Chatterjee, 2007). Granger causality in time series data was mathematically formalized by Granger (1969) and later implemented to a multivariate framework by Geweke (1984), allowing the identification of conditional relationships.

Recently, Fujita *et al*. (2010) proposed the application of Granger causality to identify relationships between *sets* of genes, i.e. to identify if a set containing n genes Granger causes another set with m genes. Their main goal is to create networks representing pathway-level connections that could help the understanding of molecular mechanisms underlying biological processes. In the above-mentioned work, the authors propose a bootstrap procedure to test and identify Granger causality. Despite the usefulness of the bootstrap method to identify/test Granger causality, it is computationally intensive, becoming particularly slow when accurate P-values are required.

To overcome this drawback, here we propose an analytical approach based on likelihood ratio test (LRT). For time series experiments containing less than 100 points, we also propose a Bartlett correction in order to improve the asymptotic approximation of the statistics.

Two simulations were carried out in order to evaluate the performance of the proposed LRT with Bartlett correction-based test. First, since the Granger causality between sets of gene expression time series data is a generalization of the multivariate version proposed by Geweke (1984), it is necessary to know if the proposed LRT is as good as the classic Wald's test in terms of statistical power in a multivariate context by using the vector autoregressive (VAR) model (Simulation 1). The second simulation consists in comparing the performance of the proposed test with the previously published bootstrap-based method (Fujita *et al*., 2010) in a context where one would like to identify Granger causality between sets of time series data.

The results show that: (i) the LRT is equivalent to Wald's test in the multivariate model in terms of statistical power; (ii) the LRT is much faster and has a higher statistical power than the bootstrap method (when analyzing sets of time series data); and (iii) both the LRT and bootstrap can control the rate of false positives even under non-Normal noises. An R package named gGranger containing the identification method for Granger causalities, the bootstrap procedure and the LRT with Bartlett correction is provided.

2 METHODS

2.1 Granger causality for sets of time series

For the linear case of Granger causality between sets of genes, let $\mathbf{Y}_t^i$ (m-dimensional) and $\mathbf{Y}_t^j$ (n-dimensional) be two disjoint subsets of $\mathbf{Y}_t$, where $\mathbf{Y}_t$ is a k-dimensional set of stationary gene expression time series ($k \geq m+n$) with length T. Then, in an autoregressive process of order one, $\mathbf{Y}_t^j$ is Granger non-causal for $\mathbf{Y}_t^i$ partialized by $\mathbf{X} = \mathbf{Y}_t \setminus \mathbf{Y}_{t-1}^j$ if the following condition holds (Fujita *et al*., 2010):

$$\mathrm{CCA}(\mathbf{Y}_t^i, \mathbf{Y}_{t-1}^j | \mathbf{X}) = 0 \quad (1)$$

where CCA is the canonical correlation analysis.

*To whom correspondence should be addressed.

2.2 LRT

Testing Equation (1) is equivalent to test $\mathbf{a}'\Sigma_{\mathbf{Y}_t^i\mathbf{Y}_{t-1}^j|\mathbf{X}}\mathbf{b}=0$, where $\Sigma_{\mathbf{Y}_t^i\mathbf{Y}_{t-1}^j|\mathbf{X}}$ is the covariance between $\mathbf{Y}_t^i$ and $\mathbf{Y}_{t-1}^j$ partialized by $\mathbf{X}$ and $\mathbf{a}$ and $\mathbf{b}$ are the coefficient vectors that maximize the correlation between $\mathbf{a}'\mathbf{Y}_t^i$ and $\mathbf{b}'\mathbf{Y}_{t-1}^j$ given $\mathbf{X}$. Therefore, the Granger causality test for sets of time series experiments may be set as $H_0:\Sigma_{\mathbf{Y}_t^i\mathbf{Y}_{t-1}^j|\mathbf{X}}=\mathbf{0}$ (Granger non-causality) versus $H_1:\Sigma_{\mathbf{Y}_t^i\mathbf{Y}_{t-1}^j|\mathbf{X}}\neq\mathbf{0}$ (Granger causality) for the following statistics:

$$-2\ln\Lambda=(T-1-(k-n))\ln(|\hat{\Sigma}_{\mathbf{Y}_t^i\mathbf{Y}_t^i|\mathbf{X}}||\hat{\Sigma}_{\mathbf{Y}_{t-1}^j\mathbf{Y}_{t-1}^j|\mathbf{X}}|/|\hat{\Sigma}|)$$

where $\hat{\mathbf{\Sigma}}=\begin{pmatrix}\hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i\mathbf{Y}_t^i|\mathbf{X}} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_t^i\mathbf{Y}_{t-1}^j|\mathbf{X}}\\ \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j\mathbf{Y}_t^i|\mathbf{X}} & \hat{\mathbf{\Sigma}}_{\mathbf{Y}_{t-1}^j\mathbf{Y}_{t-1}^j|\mathbf{X}}\end{pmatrix}$ is the block matrix of the unbiased estimator of Σ (Johnson and Wichern, 2002). For large T, the statistic test is approximately distributed as a χ^2 random variable.

For relatively small T, we suggest the use of Bartlett correction (Bartlett, 1939). Timm and Carlson (1976) showed that the Bartlett correction under our problem arises by replacing the multiplicative factor $T-1-(k-n)$ in the likelihood ratio statistic with the factor $r=T-2-(k-n)-\frac{1}{2}(m+n+1)$ to improve the χ^2 approximation to the sampling distribution of $-2\ln\Lambda$. Thus, we reject H_0 at significance level α if

$$r\ln(|\hat{\Sigma}_{\mathbf{Y}_t^i\mathbf{Y}_t^i|\mathbf{X}}||\hat{\Sigma}_{\mathbf{Y}_{t-1}^j\mathbf{Y}_{t-1}^j|\mathbf{X}}|/|\hat{\Sigma}|)>\chi^2_{mn}(\alpha)$$

where $\chi^2_{mn}(\alpha)$ is the upper (100α)th percentile of a χ^2 distribution with mn degrees of freedom.

2.3 Experimental set-up

Two simulations were carried out in order to evaluate the performance of the proposed LRT with Bartlett correction.

2.3.1 Simulation 1—Multivariate model Description of the comparison between the proposed approach and the multivariate VAR model with Wald's test in the case where $n=9$ time series (predictors) Granger cause $m=1$ time series (target) (Supplementary Fig. 1).

$$\begin{cases} y_t &= -0.1y_{t-1}+0.2x_{1,t-1}-0.3x_{2,t-1}+0.4x_{3,t-1}\\ &\quad -0.5x_{4,t-1}+0.6x_{5,t-1}-0.7x_{6,t-1}+0.8x_{7,t-1}\\ &\quad -0.9x_{8,t-1}+0x_{9,t-1}+\varepsilon_t\\ x_{i,t} &= 0.2x_{i,t-1}+\varepsilon_{i,t}\end{cases}$$

for $t=1,\ldots,T$, where the noises $\varepsilon_{i,t}\sim N(0,1)$ $(i=1,\ldots,9)$. The time series length is set to 75, i.e. $T=75$.

2.3.2 Simulation 2—Module–module model Comparison between the proposed LRT and the bootstrap-based test proposed by Fujita *et al*. (2010). The structure of the artificial network is similar to Simulation 1 described in Fujita *et al*. (2010), where the Granger causality from one gene set to another gene set is tested (Supplementary Fig. 2). To evaluate the robustness of LRT under different noise distributions, gene time series are generated with residues following both Normal and non-Normal distributions (Exponential, Uniform, Gamma, half-Normal and *t*-Student). The time series length is set to $T=75$.

3 RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed likelihood ratio-based test to identify Granger causality in both a multivariate condition and between sets of time series data, two simulations were carried out (see Section 2.3).

Figure 1a and b show the ROC curves of LRT versus Wald's test in Simulation 1 and the LRT versus bootstrap-based test in Simulation 2, respectively. The ROC curves show that the LRT and Wald's test are equivalent in the multivariate model (the solid and dashed lines coincide), and the LRT performs better than the bootstrap procedure in terms of statistical power in the module–module model (the solid line is above the dashed line). The fact that the LRT and Wald's test are equivalent in the multivariate model is an expected result since the identification of Granger causality between sets of genes in time series experiments is a generalization of the multivariate Granger causality proposed by Geweke (1984).

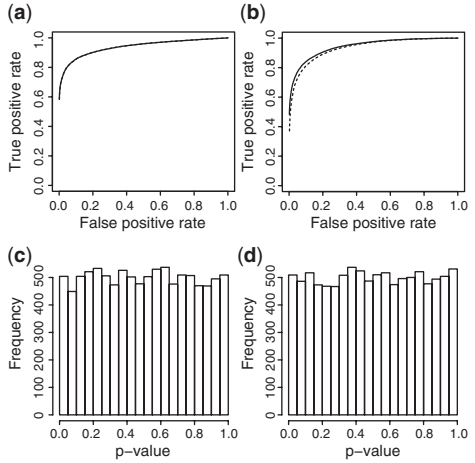

Fig. 1. (**a**) and (**b**): ROC curves obtained by running Simulations 1 and 2, respectively. Solid line represents the LRT. Dashed line is the Wald's test of the VAR model. (**c**) and (**d**): LRT *P*-values histograms under the null hypothesis obtained by running Simulations 1 and 2, respectively. LRT and Wald's test are equivalent in the multivariate model. In the module–module model, LRT performs better than the bootstrap approach.

Figure 1c and d illustrate the *P*-value distributions under the null hypothesis obtained by running 10 000 times Simulations 1 and 2, respectively, with Normal residues. Notice that the rate of false positives is effectively controlled since the histograms are close to Uniform distributions.

Simulation 2 was also performed with residues following non-Normal distributions namely, (i) Exponential, (ii) Uniform, (iii) Gamma, (iv) half-Normal and (v) *t*-Student in order to evaluate the robustness of the tests. LRT was able to control the rate of false positives in the majority of the non-Normal residues even when the residues followed *t*-Student distribution with three degrees of freedom (i.e. time series with outliers due to heavy tails in the *t*-Student distribution if compared to Normal distribution). In the case of noise following multivariate *t*-Student distribution (when there are outliers and time series are contemporaneously correlated), the Type I error was controlled but not as well when compared to the simulations performed with other noise distributions. In the multinormal distribution case (contemporaneously correlated time series without outliers), both bootstrap and LRT controlled the rate of false positives (refer to Supplementary Section 'Simulation 2 with different noise distribution' for more information about *P*-values' distributions and ROC curves in these simulations).

To verify how Type I error is controlled in actual biological data when noise distribution is unknown, we used the dataset of genes originally published by Whitfield *et al*. (2000) and used to construct networks in Fujita *et al*. (2010). The points of the time series were permuted to eliminate any presence of Granger causality. LRT was

Table 1. LRT and bootstrap procedures (with 1000, 5000 and 10 000 bootstrap samples) processing time in 1000 simulations

	Minimum	25% quantile	Median	75% quantile	Maximum
LRT	0.039	0.046	**0.048**	0.050	0.079
1000	36.586	41.698	**41.802**	41.937	67.949
5000	248.525	249.691	**250.132**	250.723	397.618
10 000	347.584	414.948	**415.991**	416.961	708.712

All processing times were calculated by using a Intel Quad Core Xeon E5450 3.0 GHz. The results are presented in seconds.

applied to this permuted time series and the process was re-done 10 000 times. The results (Supplementary Section 'Verifying the control of Type I error in actual biological data') show *P*-values histograms close to Uniform distributions under the null hypothesis, confirming that even in real biological data, LRT is able to control the rate of false positives. In addition, the method we propose here was applied to the same dataset (Whitfield *et al*., 2000) in order to exemplify its application in the context of microarray data interpretation (Supplementary Section 'Application to biological data').

For comparative purposes, the processing times of both the LRT and the bootstrap test were measured by running 1000 times Simulation 2. Table 1 shows that as the number of bootstrap samples increases in order to improve the *P*-value estimation, it becomes slower than the LRT. A fast statistical test such as the proposed LRT is of extreme importance nowadays due to the enormous amount of data generated by microarray experiments and, consequently, the increasing number of hypothesis to be tested. Bonferroni or FDR correction are usually used for multiple tests correction and when using bootstrap approaches, more accurate *P*-values must be calculated by increasing the number of bootstrap samples.

In summary, the advantages of the LRT with Bartlett correction over the bootstrap approach are: (i) the statistical power of the LRT is greater than the bootstrap procedure; (ii) the LRT can control the rate of false positives on both Normal and even in non-Normal (Exponential, Uniform, Gamma, half-Normal and *t*-Student) residues; and (iii) the LRT is much faster than the bootstrap test. As a final consideration, despite the advantages presented for LRT, we recommend the use of the bootstrap procedure, which is non-parametric, if one cannot guarantee that the residues follow one of the studied distributions.

ACKNOWLEDGEMENTS

Computational time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

Funding: Grant of RIKEN, Japan; Grant of FAPESP, Brazil.

Conflict of Interest: none declared.

REFERENCES

Bartlett,M.S. (1939) A note on tests of significance in multivariate regression. *Proc. Camb. Phil. Soc.*, **35**, 180–185.

Fujita,A. *et al*. (2009) The impact of measurement errors in the identification of regulatory networks. *BMC Bioinformatics*, **10**, 412.

Fujita,A. *et al*. (2010) Identification and quantification of Granger causality between gene sets. *J. Bioinform. Comput. Biol.*, **8**, DOI:10.1142/S0219720010004860.

Geweke,J. (1984) Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.*, **79**, 907–915.

Granger,C.W.J. (1969) Investigating causal relationships by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Guo,S. *et al*. (2008) Uncovering interactions in the frequency domain. *PLoS Comput. Biol.*, **4**, e1000087.

Johnson,R.A. and Wichern,D.W. (2002) *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey.

Mukhopadhyay,N.D. and Chatterjee,S. (2007) Causality and pathway search in microarray time series experiment. *Bioinformatics*, **23**, 442–449.

Timm,N.H. and Carlson,J.E. (1976) Part and bipartial canonical correlation analysis. *Psychometrika*, **41**, 159–176.

Whitfield,M.L. *et al*. (2000) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.