

LSA: A web service for local similarity analysis of time sequence data

Li Xia¹, Jed A. Fuhrman², Fengzhu Sun³

1 Introduction

Recent progress in high-throughput experimental technologies, especially the availability of collective spatial and temporal measurement of gene transcription, translation and microbial community activity, has generated a large number of temporal/spatial series datasets: GEO, Array Express, and Microbial Ecology Data ([1]). These datasets typically composed of temporal and/or spatial series of tens to thousands of variable. One fundamental pursuit underlying the analysis of these datasets is to identify the correlated patterns between variables/factors. Traditional correlation analysis methods, like Principle Component Analysis(PCA) , Multidimensional Scaling (MDS) , Canonical Correlation Analysis(CCA) , are typically used. However, they cannot identify the correlation due to time delays or correlation within only a limited time interval.

The Local Similarity Analysis(LSA) technique ([2], [3]) has been proposed to overcome these problems. LSA has two advantages over other approaches. First, it can identify time/space shifted correlations among the variables. Second, it can identify correlated pairs even if they are correlated under a limited temporal or spatial region. In addition, a directed correlation network of original variables can potentially be inferred if lagged association is identified. The LSA has already shown to be a useful tool in the analysis of gene expression analysis ([3]) and microbial ecology data ([2]). We believe that an easy-to-use LSA package and web server will help the scientists use the technique in many other applications.

Therefore, the LSA package described in this paper is aimed to provide a user-friendly toolkit for the general scientific community. The LSA package is implemented in Python Programming language (<http://www.python.org>). It takes advantage of Python's efficient numerical models Numpy and Scipy (<http://www.scipy.org>) and is ported as a standard Python module. The web server provides an easy to use interface for LSA, meanwhile the software provides easy source code level access to executable, libraries and documentation for advanced users. The analysis result is available as plain text table for direct access. It can also be processed into a SIF formatted interaction network file for visualization and analysis in other software packages, such as Cytoscape.

We use a marine ecology dataset for a demonstration of LSA (see Table 1. This dataset is the abundance of 176 bacteria types measured at 35 time points (roughly monthly over three years). In the whole pipeline, our LSA server successfully captured those time-shifted correlations (Table 1), built a correlation network, and produced the co-varying graph for interested variable pairs. This software package can be used to analyze other time series data including gene expression profiles.

2 Software and Files

web service: <http://meta.usc.edu/lsa>

software: <http://128.125.86.98/svn/repos/lsa/tags/current/>

3 Figures and Tables

¹Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, RRI 201, Los Angeles, CA 90089-2910 USA E-mail: lxia@usc.edu

²Department of Biological Sciences, University of Southern California, 3616 Trousdale Pkwy, AHF 107, Los Angeles, CA 90089-0371 USA, E-mail: fuhrman@usc.edu

³Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs Way, RRI 201, Los Angeles, CA 90089-2910 USA, E-mail: fsun@usc.edu

X	Y	LS	X ₀	Y ₀	L	D	pV	Corr	C
473	593	0.6436	4	1	32	3	0.000	-0.0642	0.35504
473	768	0.5940	4	2	32	2	0.000	0.1030	0.27495
478	oxy	0.6076	1	3	33	-2	0.001	0.1714	0.15872
489	723	0.7030	1	3	33	-2	0.000	0.1947	0.12755
573	658	0.5739	1	4	32	-3	0.000	0.1797	0.14718
582	658	0.6127	2	1	34	1	0.000	0.1778	0.14976
593	775	0.6028	1	3	33	-2	0.000	0.1927	0.13014
658	677	-0.5879	3	1	33	2	0.000	0.1254	0.23310
658	855	0.6578	3	1	33	2	0.000	-0.0067	0.48443
658	1129	0.5997	3	1	33	2	0.001	0.1400	0.20770
677	723	-0.6013	1	3	33	-2	0.001	-0.1552	0.18304
677	755	-0.5926	2	4	32	-2	0.001	-0.0198	0.45431
677	775	-0.6105	1	3	33	-2	0.001	-0.0034	0.49211
723	oxy	0.5953	4	1	32	3	0.001	-0.1312	0.22289
739	858	0.5905	2	1	34	1	0.001	0.1224	0.23849
755	855	0.5985	4	2	32	2	0.001	0.1661	0.16650
755	1129	0.6614	4	2	32	2	0.000	0.1944	0.12789
775	855	0.6236	3	1	33	2	0.000	0.1424	0.20377
855	No ₃	0.6662	1	4	23	-3	0.000	0.0669	0.34905

Table 1: Example of LSA result table. The columns from left to right are: Labels of sequences X and Y, normalized Local Similarity score (LS, from -1 to +1), alignment start position in X and Y (X_0 and Y_0), alignment length (L), delay in time unit (D), P-value of LS score (pV), Pearson Correlation (Corr), P-value of Correlation (CpV). The table demonstrates LSA’s ability to capture time-shifted interactions with high LS score and pV, which would be otherwise ignored by the ordinary correlation analysis because of low Cor and high CpV.

4 References and Bibliography.

References

- [1] Fuhrman JA, Hewson I, Schwalbach S M, Steele JA, Brown MB, Naeem S 2006. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences* 103(35):13104–13109.
- [2] Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F 2006. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* 22(20):2532–2538.
- [3] Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M 2001. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology* 314(5):1053–1066.