Extended Local Similarity Analysis (eLSA) of Biological Data

Synonyms
Local similarity analysis; Local association analysis.

Introduction
The advances in high-throughput low-cost experimental technologies have made possible time-series studies of hundreds or thousands biological factors simultaneously. The availability of such datasets lead to an increased interest in profile similarity analysis techniques that can identify significant association patterns possibly embracing biological insights. In the context of metagenomics, factors of particular interest are operational taxonomic units (OTUs), microbial genomes and environmental genes. Their association patterns may suggest microbe-environment, symbiotic relationships and other types of interactions.

Many computational or statistical approaches exist to study the profile similarity at global scale, such as Pearson’s Correlation Coefficients (PCC), Spearman’s Correlation Coefficients (SCC), principal component analysis (PCA), multi-dimensional scaling (MDS), discriminant function analysis (DFA) and canonical correlation analysis (CCA). However, in many biological settings, the interaction may be active within only certain subintervals or the response to regulation may be time lagged. Methods based on the global relationships of profiles, may fail to detect these interactions.

Extended local similarity analysis (eLSA) method is specifically developed to capture local and potentially time-delayed co-occurrence and association patterns in time series data that cannot otherwise be identified by ordinary correlation analysis.

Description
Local association with possible time delays
Local association refers to the association that only occurs in a subinterval of the time of interest. Time-delayed association indicates that there is a time lag for the response of one factor to the change in another factor. As an example of local association, in Fig 1, the top-left panel shows two series X and Y with non-significant correlation (r=0.26, \(P=0.273\)); however, they are in fact significantly correlated in the time interval from 7 to 16 as shown in the bottom-left panel (eLS=0.43, \(P=0.028\)). As an example of time-delayed local association, in Fig 1, the top-right panel shows two series X and Y with non-significant correlation (r=-0.26, \(P=0.272\)); however, they are in fact significantly correlated in time interval from 4 to 17 if X is shifted three units toward origin as shown in the bottom-right panel (eLS=0.51, \(P=0.006\)).
Extended local similarity analysis

Extended local similarity analysis (eLSA) is an analysis technique designed to capture local associations possibly with time-delays. eLSA extends the original local similarity analysis technique (Qian et al. 2001; Ruan et al. 2006) and local shape analysis technique to time series data with replicates (Xia et al. 2011). Improvements in computation efficiency of p-values are also made (Xia et al. 2012). Time series data of a pair of factors X and Y with replicates can be expressed as data matrices $X_{[1:m][1:n]}$ and $Y_{[1:m][1:n]}$, where each column is one sample from the time point and n is the number of time points; each row is a replicate and m is the number of replicates.

Given time series data of two factors and a user-constrained delay limit, eLSA uses dynamic programming algorithm to find the configuration of the data that yields the highest extended local similarity (eLS) score - a similarity metric defined as:

$$|eLS(X_{[1:m][1:n]}, Y_{[1:m][1:n]})| = \frac{1}{n} \max_{i,j,l \text{ s.t. } |i-j| \leq D} \sum_{k=0}^{l-1} F(X_{[1:m],i+k}) F(Y_{[1:m],j+k})$$

where D is the delay limit and F is the summarizing function for repeated measures (mean, median, etc.). For example, within a delay limit of two units, the first time spot of one series might be aligned to the third time spot of the other series, thus maximizing their eLS.

For a dataset of many factors, eLSA is applied to each pairwise combination of factors in the dataset. Candidate associations are then evaluated statistically by a permutation test, which calculates the p-value - the proportion of scores exceeding the original eLS score after shuffling the first series and re-evaluating the eLS score many times; or more efficiently by theoretical approximation. Researchers can use eLSA to detect undirected associations, i.e., association patterns without time delays, and directed associations, where the change of one factor may temporally lead or follow another factor. Fig 2 shows the analysis pipeline of the eLSA technique.
Fig 2. The eLSA pipeline. Users start with raw data (matrices of time series) as input and specify their requirements as parameters. The LSA tools subsequently F-transform and normalize the raw data and calculate extended Local Similarity (eLS) scores and Pearson’s Correlation Coefficients. The tools then assess the statistical significance (P-values) of these correlation statistics using the permutation test and filter out insignificant results. Finally, the tools construct a partially directed association network from the significant associations.

Inferring co-occurrence networks using eLSA

Studies adopting the local similarity analysis technique have shown interesting and novel discoveries for microbial community network analysis. In one of the studies (Steele et al. 2011), eLSA is used to find associations among relative abundances of bacteria, archaea, protists, total abundance of bacteria and viruses, and physico-chemical parameters. Co-occurrence networks were generated from significant eLSA associations to visualize and identify time-dependent relationship among ecologically important taxa, for example, the SAR11 cluster, stramenopiles, alveolates, cyanobacteria and ammonia-oxidizing archaea.

A subnetwork from the study is shown in Fig 3. It is built around γ-proteobacteria OTUs as central nodes (abbreviated Alt: alteromonas, CHB: CHABI-7, Gam: γ-proteobacterium, S86: SAR86, S92: SAR92). This subnetwork identifies 12 γ-proteobacterial OTUs. γ-proteobacteria OTUs correlate with eukaryotes and Crenarchaea (Cren), as well as environmental parameters and bacterial production. γ-proteobacterium SAR92-749 is more likely opportunistic species, as the relative abundance of SAR92-749 positively correlated with bacterial production measured by leucine and thymidine incorporation (eLS=0.54, P=0.003 and eLS=0.495, P=0.005, respectively).
Fig 3. An eLSA subnetwork built around γ-proteobacteria OTUs as central nodes (abbreviated Alt: alteromonas, CHB: CHABI-7, Gam: γ-proteobacterium, S86: SAR86, S92: SAR92).

Conclusion

eLSA technique uniquely captures local and potentially time-delayed co-occurrence and association patterns in time series data. eLSA technique is also applicable to other types of gradient data, including the response to different levels of treatments, temperature, humidity, or spatial distributions. The analysis pipeline is implemented as a C++ extension to Python, which streamlines data normalization, local similarity scoring, permutation testing and network construction. More information about the software is available from eLSA’s homepage at http://meta.usc.edu/softs/lsa.

Cross references

Understanding Composition and Dynamics of Microbial Communities, New Tools, Project
Computational Approaches for Metagenomic Datasets
Marine bacterial, archaeal and protistan association networks
Marine bacterioplankton species and environmental factors: Local similarity analysis
Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads

References

Extended Local Similarity Analysis (eLSA) of Biological Data

Professor Fengzhu Sun
Department of Biological Sciences, University of Southern California, Los Angeles, USA

PhD Li Charlie Xia
Computational Biology and Bioinformatics, University of Southern California, Los Angeles, United States

DOI: 10.1007/SpringerReference_332224
URL: http://www.springerreference.com/index/chapterdbid/332224
Part of: Encyclopedia of Metagenomics
Editor: Dr. Karen E. Nelson
PDF created on: October, 03, 2012 00:09

© Springer-Verlag Berlin Heidelberg 2012