1.研究背景

基因彼此之间的作用、环境学中：理解不同有机物以及环境因素，比如：食物来源、温度等等，彼此作用进而影响整体生态；(Understanding how genes regulate each other and when the

regulations are active is an important problem in molecular biological research. Similarly, in ecological studies, it is important to understand how different organisms and environmental factors, such as food resources, temperature, etc., regulate each other to affect the whole community. )

时间序列数据可以洞察这些问题；(Time series data can give significant insights about the regulatory relationships among different factors.)

绝大多数的方法是对整体时间进行研究的；(.Most of these methods consider the correl

ation of expression patterns across the entire time interval of interest.)

然而在许多研究中，我们需要探究时间间隔之间的信息；(For many gene regulation relationships, the regulation may be active in certain subintervals. Methods based on the

global associations of the gene expression profiles may fail to detect these relationships.)

2.研究现状

许多局部联系的方式已经被开发出来进而解决这些问题；(Several local association-based methods have been developed to address this problem (Qian et al., 2001; Balasubramaniyan et al., 2005; Ji and Tan, 2005; He and Zeng, 2006).)

Qian等人借用了来自于分子序列的局部分布，提出去识别在时间偏移的意义下的局部相似性(Borrowing the idea from local alignment for molecular sequences, Qian et al. (2001) proposed to identify local and potential time-delayed (- lagged) associations between gene expression profiles.)

局部就是表明两个因素仅仅在一些时间间隔的联系，而时间偏移却是指这个两个变量之间的超前以及滞后；(Here, local indicates the two factors are only associated within some time subinterval, and time-delayed indicates there is time shift in the associated profiles.)

局部相似性是被LS变量进行测量；(The strength of local association is measured by local similarity (LS) score and the statistical significance of LS score is evaluated by a large number of permutations.)

作者展示这种分析完全可以识别全局探测不能得出的局部相似时间间隔对；

Ruan等人使用局部相似性分析去研究超过四年在海洋中的微生物之间的联系；

而且这个方法已经在许多最近的生态学研究之中被应用；(The authors showed that such analysis can identify associated pairs that are not detectable through global analysis. Ruan et al. (2006) used a similar approach to study local associations of microbial organisms in the ocean over a 4-year period, and this approach has been used in several other recent ecological studies (Beman et al., 2011; Chaffron et al., 2010; Gilbert et al., 2011; Shade et al., 2010; Steele et al., 2011).)

Charli xia等人最近将这种方法进行了拓展去解决多次测量数据的问题，并且不仅仅获得了分析数据，还有置信区间；(Xia et al. (2011b) recently extended the approach to deal with replicated time series where not only statistical significance of LS score can be evaluated, but also a bootstrap confidence interval can be obtained.)

3.研究问题

一个关于局部相似性分析的主要局限就在于permutation的时间耗费，当有G条时间序列数据时，就需要G(G-1)/2的对进行分析；(One of the major limitations of the local similarity analysis is the time-consuming permutation procedure used to evaluate the statistical significance (P-value) of the LS score. When a large number of (G) genes are considered, GðG 1Þ=2 gene pairs need to be

evaluated. )

4.研究成果

我们开发出来有关于局部相似性分析的理论方法可以作为局部相似性分析P值计算方法的替代品；

这种方法的时间消耗总是恒定的，降低了在大规模比对分析中的计算负担；

比如在宏基因组分析中，在短暂的读取文件以及丰富的估计计算，就可以得到成千上万的微生物相似性分析；(The theoretical statistical significance approximation we proposed in this work can serve as an efficient alternative for calculating P-values in local similarity analysis. Its time cost is always constant, which reduces the computational burden in a large scale pairwise analysis. For example, in metagenomics, after short read assignment and abundance estimation (Xia et al., 2011a; He and Xia, 2007), profiles of thousands of microbial OTUs are available.)

在这项工作之前，在这个数量级的局部相似性分析使用permutation是十分艰难的；

如果使用permutaition，就必须进行并行计算以及使用硬件进行加速，这就增加了分析计算的困难程度；

但是有了这个新方法之后，研究人员就可以使用笔记本电脑进行快速计算，允许飞速的网络挖掘以及分析；(Before this work, pairwise local association analysis with this number of factors was hardly tractable using permutation procedures, if not impossible. Parallel computation and hardware acceleration or some pre-clustering and filtering approaches were required, increasing the difficulty of analysis. With the new method, researchers can quickly compute the statistical significance for all OTU pairs on desktop computers, allowing on-the-fly network mining and analysis.)

5.评价

这篇文章的组织如下所述。在“方法”部分，作者提供了推导LS分数高于阈值的近似尾部概率的理论基础。

在“结果”部分，作者使用模拟研究了理论近似成立所需的数据点数量n。还使用理论公式研究了来自不同高通量实验的三个真实数据集：microarray, molecular finger printing and NGS tag-sequencing（微阵列、分子指纹和NGS标记测序）。

在“结论”部分，文章以对进一步应用和未来研究方向的讨论作为结束。