RNA-seq共表达网络的典型相关性分析

摘要

由下一代测序技术支持的数字化转录组分析发现了大量的mRNA变体。基因表达的多样化是很多生物过程的基础，同时也是阐明一些常见疾病机制的关键所在。然而，现有的基于全基因表达的构建共表达网络的方法都是针对于微阵列表达数据的。它们忽略了大量基因表达过程中的变体。为了利用exon、基因位点层面、等位基因表达的信息，我们开发了新的基于组件（component-based）的方法。分别是一元以及二元的 基于RNA-seq数据构建共表达网络的 典型相关性分析。为了评价基于RNA-seq数据的共表达网络推断（co-net inference），这两个方法应用在了TCGA-LUSC表达数据（Lung squamous cell cancer，肺鳞状细胞癌）、双极无序精神分裂症的RNA-seq数据。初步的结果显示，我们的方法建立的共表达网络提供了丰富的遗传和分子层面的信息，有助于深入洞察一些生物过程和疾病机理。我们提出的新方法，大体上优于现有的其它的统计方法（在全基因表达水平上基于RNA-seq或微阵列数据的共表达网络构建）。

介绍

不管复杂疾病的基因研究所获取的重大成就，已被识别的基因变体相关功能的研究仍然受到很多限制。基因表达变体可能对表现型的差异有重要的影响。基因表达的分析是研究基因表达变体的重要的原始资料，并且在揭示一些复杂特性的机制方面扮演着重要的角色。目前快速发展的下一代测序技术已经成为基因表达图谱构建平台的一种选择。基因表达图谱的RNA-seq提供了一个综合而广泛的转录组全景，并且比微阵列平台更加出众。RNA-seq在基因表达分析方面做出了一系列定性以及定量的改进，提供了多层解析以及转录组复杂度：exon表达、单核多态（SNP，single-nucleotide plymorphism），位点水平（positional level）、剪切、整个基因上的转录后RNA编辑、isoform以及等位基因表达（allele-specific expressions，ASE）。

变异在复杂的表型中并不是有单个基因作为一个标记者导致的，而是由一组交互的基因（通常组成一个多钟类型的生物网络）。可以从基因共表达网络中提取一些重要信息about一组组共调控基因，这些基因在调控过程中扮演着重要的角色。共表达网络可以从大体上捕捉被环境扰乱的 转录组的单独成分之间的交互关系。因此，这是强有力的工具，用来深入洞悉基因功能、生物过程、转录组的全局结构、复杂疾病的机理。

传统的构建共表达网络的统计学方法，如weighted co-expression networks、mutual information relevance networks、covariance selection、sparse graphical model，partial correlation methods，这些方法都是为微阵列数据而设计的。这些方法都是用单值统计量来带包基因表达水平，却忽略了所有信息on exons、genomic position、等位基因的差异表达。因此，即使RNA-seq大幅度地细化了生物学的研究层次，但我们仍在使用传统的统计学方法来进行共表达网络的推断（inference），这些方法是专门为微阵列数据设计的，并没有有效利用RNA-seq中包含的大量信息。为了有效利用这些信息provided by RNA-seq，强有力的计算工具亟待开发。

在这片文章里，我们开发了一种计算方法，用来解决利用RNA-seq数据进行共表达网络的Inference。为了探索exons、genomic position的表达变体across基因，我们使用了一种平凡单变量CCA来定量表示两个exon表达量（或position level表达量）的线性组合in one gene，来构建共表达网络。Exon表达量（或genomic position）将作为变量。两个基因的这种表达量分别组成了两个大的变量集。我们想重点研究相互之间具有较大关联性的线性组合。CCA的目的就是寻找那些能最大化两组变量之间关联性的线性组合。为了达到这个目的，第一步先找出关联性最强的两个线性组合。接着再找出和之前选出的那对没有关系但相互之间关联性最大的两个线性组合。接着重复上述步骤。因此CCA同样可以测量两个基因之间genomic postion或者ASE的共表达。为了建立ASE共表达网络模型，我们开发了二元CCA算法。二元CCA测量的是两个向量集合之间的关联性。一个实现CCA for 共表达网络构建的算法的程序可以再[www.bioconductor.org](http://www.bioconductor.org)或者<http://www.sph.uth.tmc.edu/hgc/faculty/xiong/index.htm>上下到。

材料和方法

数据获取：TCGA LUSC，KEGG pathways。

CCA-构建基因共表达网络