组会汇报

汇报人：Lilian

2022/1/13

文献来源

nature methods

Published: 26 January 2014

Similarity network fusion for aggregating data types on a genomic scale

Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains & Anna Goldenberg

用于在基因组规模上聚合数据类型的相似网络融合

内容概述

- ✓ 最近的技术使收集各种类型的全基因组数据具有成本效益。需要计算方法来组合这些数据，以创建给定疾病或生物过程的综合视图。
- ✓ 相似性网络融合 (SNF) 通过为每种可用数据类型构建样本（例如患者）网络，然后将这些网络有效地融合到一个代表所有基础数据的网络中，从而解决了这一问题。
- ✓ 例如，为了在给定一组患者的情况下创建疾病的综合视图，SNF 计算并融合从每个数据类型分别获得的患者相似性网络，充分**利用了数据的互补性**。
- ✓ 作者使用 SNF 结合了五个癌症数据集的 mRNA 表达、DNA 甲基化和 microRNA (miRNA) 表达数据。

Introduction

- ✓ 快速发展的技术使收集多种多样的基因组规模数据集来解决临床和生物学问题变得越来越容易。例如，癌症基因组图谱(TCGA)的大规模工作已经收集了来自数千名患者的20多种癌症的基因组、转录组和表观基因组信息。如此丰富的数据**使得综合方法对于捕捉生物过程和表型的异质性至关重要**，例如，导致乳腺癌中同源亚型的识别。
- ✓ 数据集成方法需要克服至少三个计算挑战:

(i) 样本数量较少;

(ii)每个数据集在规模、收集偏差和噪声方面的差异

(iii)不同类型数据所提供的信息的互补性。

目前的整合方法还不能同时解决所有这些问题

Introduction

先前研究的不足

- ✓ 基于拼接的方式：将不同生物学域(如mRNA表达和DNA甲基化)的标准化测量数据串联起来。

  存在低信噪比的问题；

- ✓ 独立分析后拼接：在组合数据之前分别分析每种数据类型

  数据独立，容易产生不同的输出结果，不能统一；

- ✓ 基因预先选择：从每个数据源中预先选择一组重要基因，并使用共识聚类(Consensus Clustering1)对数据进行合并。

  聚焦公共信息，缺失了互补信息；

- ✓ 机器学习聚类——iCluster：使用了一个联合潜在变量模型来进行整合聚类

  对于预先选择的基因数量特别敏感。

Introduction

本文提出SNF：Similarity network fusion：

- 为每一类构建一个**相似度网络**；
- 用**非线性方法融合**所有的相似度网络得到一个**单一的输出网络。**

SNF的优势：

- 同时包含**不同基因类型的公共信息和互补信息**，提取的信息比较全面；
- 可以**综合处理多种基因数据，对噪声鲁棒，可用于样本少的情况**；
- 迭代融合的过程可以**去除弱连接，增强强连接。**

Datasets

- ✓ 使用了TCGA网站提供的**五种**不同癌症类型的数据:**GBM、BIC、LSCC、KRCCC和COAD。**
- ✓ 对于每一种肿瘤类型，下载了TCGA管理的3级数据集，**包含基因表达、miRNA表达和DNA甲基化信息。**
- ✓ TCGA存储库为每种数据类型包含多个平台，**选择与最大数量的可用个体相对应的平台，并尽可能描述肿瘤样本和对照。**

  在GBM和LSCC中使用了Broad Institute HT-HG-U133A平台，
  在BIC和COAD中使用了UNC-Agilent-G4502A-07平台，
  在KRCCC中使用了UNC-Illumina-Hiseq-RNASeq平台。
- ✓ 对于miRNA表达数据

  在BIC中使用了BCGSC-Illumina-Hiseq-miRNAseq平台，
  在GBM中使用了UNC-miRNA-8X15K平台，
  在LSCC、KRCCC和COAD中使用了BCGSC-Illumina-GA-miRNAseq平台。
- ✓ 对于甲基化数据

  在GBM中使用了jhu - usc - illumina - dna -甲基化平台，
  在BIC、LSCC、KRCCC和COAD中使用了JHU-USC-Human-Methylation-27平台。

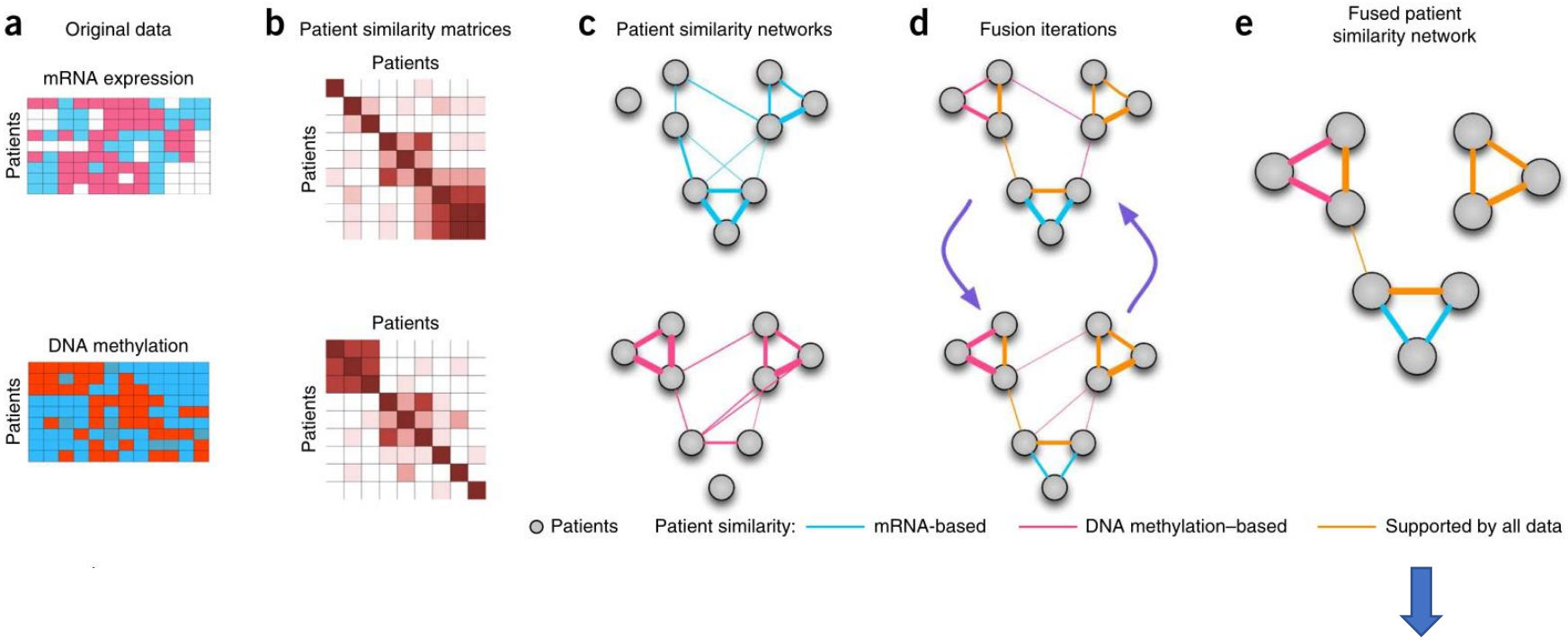
Preprocessing

在应用我们的SNF之前，执行了三个步骤的预处理:离群值去除，缺失数据查补和归一化。

如果某一患者某一数据类型的缺失数据超过20%，则不考虑该患者。

同样，如果某个生物学特征（例如，mRNA 表达）在患者中具有超过 20% 的缺失值，则过滤掉这个特征。

SNF步骤的演示示例



(a)同一队列患者的mRNA表达和DNA甲基化数据集的示例表示。
(b)每种数据类型的患者之间的**相似度矩阵**。
(c)患者-患者相似度网络，相当于患者-患者数据。**用节点表示患者，用边缘表示患者的配对相似度。**
(d) SNF网络融合用来自其他网络的信息**迭代更新每一个网络**，使每一步网络更加相似。
(e)迭代的网络融合结果收敛到最终的融合网络。**边的颜色表示哪些数据类型对给定的相似性有贡献。**

Similarity network fusion.

Similarity network fusion. Suppose we have n samples (for example, patients) and m measurements (for example, mRNA gene expression). We will use the patient network example throughout this section for clarity though the method has broad applicability as discussed above. A patient similarity network is represented as a graph $G = (V, E)$. The vertices V correspond to the patients $\{x_1, x_2, \cdots, x_n\}$ and the edges E are weighted by how similar the patients are. Edge weights are represented by an $n \times n$ similarity matrix $\mathbf{W}$ with $\mathbf{W}(i, j)$ indicating the similarity between patients x_i and x_j and are computed as follows. We denote $\rho(x_i, x_j)$ as the Euclidean distance between patients x_i and x_j. We then use a scaled exponential similarity kernel to determine the weight of the edge:

$$\mathbf{W}(i,j) = \exp\left(-\frac{\rho^2(x_i,x_j)}{\mu\varepsilon_{i,j}}\right) \quad (1)$$

where μ is a hyperparameter that can be empirically set and $\varepsilon_{i,j}$ is used to eliminate the scaling problem. Here we define

$$\varepsilon_{i,j} = \frac{\text{mean}(\rho(x_i,N_i)) + \text{mean}(\rho(x_j,N_j)) + \rho(x_i,x_j)}{3}$$

1. 边缘权值由n × n个相似度矩阵W表示，其中W(i, j)表示患者xi和xj之间的相似度
2. 把ρ(xi, xj)表示为病人xi和xj之间的欧氏距离
3. 使用比例指数相似核（scaled exponential similarity kernel）来确定边的权值

Similarity network fusion.

$\varepsilon_{i,j}$ is used to eliminate the scaling problem. Here we define

$$\varepsilon_{i,j} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}$$

where mean $(\rho(x_i N_i))$ is the average value of the distances between x_i and each of its neighbors. We recommend setting μ in the range of [0.3, 0.8]. Note that while this distance measure is suitable for continuous variables, we propose to use chi-squared distance for discrete variables and agreement-based measure for binary variables.

ε用于减小尺度问题（scaling problem），类似于一个放缩系数

Ni表示该特征下，第i个样本K个近邻的集合，为经验值，具体根据样本大小选取

Similarity network fusion.

To compute the fused matrix from multiple types of measurements, we define a full and sparse kernel on the vertex set V. The full kernel is a normalized weight matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix whose entries $\mathbf{D}(i, i) = \Sigma_j \mathbf{W}(i, j)$, so that $\Sigma_j P(i, j) = 1$. However, this normalization may suffer from numerical instability since it involves self-similarities on the diagonal entries of $\mathbf{W}$. One way to perform a better normalization is as follows:

$$\mathbf{P}(i,j) = \begin{cases} \dfrac{\mathbf{W}(i,j)}{2\Sigma_{k \neq i}\mathbf{W}(i,k)}, j \neq i \\ 1/2, j = i \end{cases} \tag{2}$$

- ✓ 表示该特征下，第i个样本与第j个样本的相似情况，
- ✓ 相比于W矩阵，做了标准化处理
- ✓ P包含了所有样本之间的相似信息，同时满足 $\Sigma_j \mathbf{P}(i,j) = 1$

Similarity network fusion.

Let N_i represent a set of x_i's neighbors including x_i in G. Given a graph, G, we use K nearest neighbors (KNN) to measure local affinity as:

$$\mathbf{S}(i,j) = \begin{cases} \dfrac{\mathbf{W}(i,j)}{\Sigma_{k \in N_i} \mathbf{W}(i,k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

- ✓ S可看作是P的KNN图，只保留了样本与其临近K个样本之间的相似信息，而距离样本较远处的信息则丢弃，置为零；
- ✓ 用S表示局部亲密度，

- ✓ 之后在不同特征下的图进行融合迭代的过程中，P作为初始状态、S作为融合迭代过程的核矩阵
- ✓ 对每一个特征都进行以上计算，完成相似网络初始化

Similarity network fusion.

Let $\mathbf{P}^{(1)}_{t=0} = \mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}_{t=0} = \mathbf{P}^{(2)}$ represent the initial two status matrices at $t = 0$. The key step of SNF is to iteratively update similarity matrix corresponding to each of the data types as follows:

$$\mathbf{P}^{(1)}_{t+1} = \mathbf{S}^{(1)} \times \mathbf{P}^{(2)}_{t} \times (\mathbf{S}^{(1)})^{T} \quad (4)$$

$$\mathbf{P}^{(2)}_{t+1} = \mathbf{S}^{(2)} \times \mathbf{P}^{(1)}_{t} \times (\mathbf{S}^{(2)})^{T} \quad (5)$$

迭代更新P的相似度矩阵 → 并行的交换扩散过程

where $\mathbf{P}^{(1)}_{t+1}$ is the status matrix of the first data type after t iterations. $\mathbf{P}^{(2)}_{t+1}$ is the similarity matrix for the second data type. This procedure updates the status matrices each time generating two parallel interchanging diffusion processes. After t steps, the overall status matrix is computed as

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}^{(1)}_{t} + \mathbf{P}^{(2)}_{t}}{2}.$$

最终的状态矩阵

Since $\mathbf{S}$ is a KNN graph of $\mathbf{P}$, which can reduce some noise between instances, our SNF is robust to the noise in similarity measures.

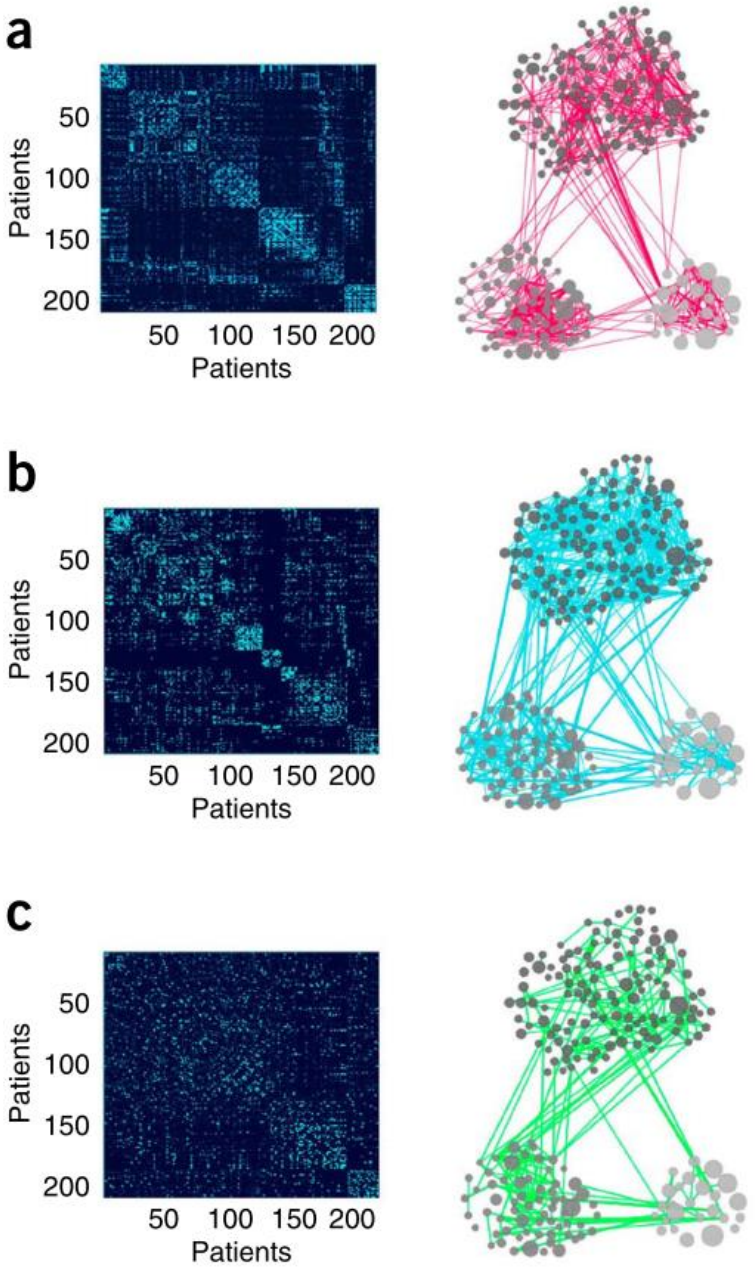
Another way to think of the updating rule (4) is

$$\mathbf{P}^{(1)}_{t+1}(i,j) = \sum_{k \in N_i} \sum_{l \in N_j} \mathbf{S}^{(1)}(i,k) \times \mathbf{S}^{(1)}(j,l) \times \mathbf{P}^{(2)}_{t}(k,l) \quad (6)$$

A case study: glioblastoma multi

- ✓ 论文以一个细胞瘤数据集进行了分析，展示了SNF的具体过程，这里使用了三种基因数据，首先也是根据这三种基因数据分别进行构图，构图后对三个图进行融合。

- ✓ DNA甲基化和mRNA表达表明相对较强的簇间相似性(图2a,b)



Conclusion

- 将患同一种癌症的病人组成一个群体，利用群体里面每个病人个体的不同基因数据分别构建不同的图，并设计了一个图融合方式，将不同的图融合成一个最终的图，最终的图包括了所有的基因信息数据，因此是一个综合的结果

- 利用该综合的结果进行聚类，可以将癌症分为不同的亚型，利用该综合结果进行回归任务，可以对病人的生存风险进行预测。

- SNF还有许多其他的应用。在临床领域，患者网络允许整合非常不同种类的测量数据，如微生物组和代谢组数据、问卷调查和功能磁共振成像，以及基因组、临床和人口统计学数据，只要这些数据可以用来识别患者之间的相似性

酵母相关数据库——ArrayExpress数据库

白色念珠菌，裂殖酵母和酿酒酵母中木糖代谢的比较。

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-GEOD-50476	Comparative xylose metabolism among the ascomycetes C. albicans, S. stipitis and S. cerevisiae	transcription profiling by array	Candida albicans SC5314, Saccharomyces cerevisiae	9	30/10/2013	⬇	⬇	-

数据描述：子囊酿酒酵母，白色念珠菌和裂殖酵母在不同程度上代谢木糖。这三个物种均包含高度相似的基因，**这些基因编码将木糖转化为木酮糖所需的木糖还原酶和木糖醇脱氢酶**...实验表明酿酒酵母具有将木糖转化为木酮糖的酶促能力。总体而言，这项工作建立了工程化的缺乏木糖代谢的必要步骤的白色念珠菌菌株，为分析多个物种的木糖代谢酶提供了平台，并证实酿酒酵母具有将木糖转化为木酮糖的遗传潜力。

Samples (9)
Less...

GSM1219874 WT in xylose (14044065)
GSM1219875 WT in xylose (14044066)
GSM1219876 WT in xylose, replica 1 (14308825)
GSM1219877 WT in xylose, replica 1, dye-swap (14
GSM1219878 WT in xylose, replica 2 (14308823)
GSM1219879 WT in xylose, replica 2, dye-swap (14
GSM1219880 WT no sugar, replica 1 (14309081)
GSM1219881 WT no sugar, replica 1, dye-swap (14
GSM1219882 WT no sugar, replica 2 (14309140)

酵母相关数据库——ArrayExpress数据库

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-GEOD-12890	Transcription profiling Saccharomyces cerevisiae xylose metabolism	transcription profiling by array	Saccharomyces cerevisiae	15	26/10/2008	⬇	⬇	-

酿酒酵母木糖代谢转录谱。

数据描述：将在木糖上好氧分批培养生长的重组酿酒酵母的转录组和蛋白质组与在葡萄糖上生长细胞进行了比较。**目的是在全基因组水平上研究在葡萄糖或木糖上生长的细胞中信号转导和碳分解代谢物阻遏有何不同。**对于进一步改造酵母以实现更有效的木糖厌氧发酵至关重要。实验总体设计：在50 g/l葡萄糖和50 g/l木糖上进行了三个有氧分批发酵，以进行比较。从木糖培养开始72小时后收获木糖生长的细胞样品，剩余32 g/1残留木糖。

酵母相关数据库——ArrayExpress数据库

Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-GEOD-835	Transcription profiling of S. cerevisiae xylose metabolism mutants grown on glucose and xylose under aerobic and anaerobic conditions	transcription profiling by array	Saccharomyces cerevisiae	6	07/07/2007		-	-

好氧和厌氧条件下在葡萄糖和木糖上生长的酿酒酵母木糖代谢突变体的转录谱。

数据描述：用于木糖代谢的酿酒酵母对碳源和通气变化的响应。**主要研究了酿酒酵母在木糖和葡萄糖代谢方面的转录调控差异，并通过实时PCR证实了关键基因的调控。**结果表明酿酒酵母在木糖上生长时需要较高的呼吸活性以及在厌氧条件下不能在木糖上生长。

酵母相关数据库——SGD数据库

Xylose-induced dynamic effects on metabolism and gene expression in engineered *Saccharomyces cerevisiae* in anaerobic glucose-xylose cultures

Susanne Alff-Tuomala, Laura Salusjärvi, Dorothee Barth, Merja Oja, Merja Penttilä, Juha-Pekka Pitkänen, Laura Ruohonen & Paula Jouhten

Applied Microbiology and Biotechnology **100**, 969–985(2016) | Cite this article

木糖对厌氧葡萄糖-木糖培养物中酿酒酵母代谢和基因表达的动态影响。

摘要：针对酿酒酵母强烈偏爱葡萄糖而不是木糖，以及两者共同消耗问题（共底物维持氧化还原平衡），以全葡萄糖底物为对照组，在整个培养过程中分析了具有XR-XDH途径利用木糖的酿酒酵母的基因表达。**模拟了基因组规模的动态通量平衡分析模型，以分析酿酒酵母的代谢动力学。**这是首次在与葡萄糖共培养的过程中，将酿酒酵母对木糖的调节和代谢反应作为时间的函数进行量化。旨在确定潜在的因素，以克服酿酒酵母对木糖的低效利用。

酵母相关数据库——GEO数据库

Genomic and phenotypic characterization of a refactored xylose-utilizing *Saccharomyces cerevisiae* strain for lignocellulosic biofuel production

Phuong Tran Nguyen Hoang, Ja Kyong Ko, Gyeongtaek Gong, Youngsoon Um & Sun-Mi Lee

Biotechnology for Biofuels **11**, Article number: 268 (2018) | Cite this article

重构的酿酒酵母菌株基因组和表型特征。

摘要：高性能木糖发酵菌株很少被用作先进的生物燃料和化学生产平台，并且需要进一步的工程设计以扩大其产品范围。在这项研究中作者重构了一种高性能的木糖发酵酿酒酵母。通过组合CRISPR–Cas9介导的合理和进化工程，**获得了新重构的基于异构酶的木糖发酵菌株XUSE，证明了木糖有效转化为乙醇的效率很高。 XUSE的基因组和转录组学分析表明，有益的突变和基因表达改变是XUSE增强木糖发酵性能的原因。**