



基于分子特征的肿瘤分类与分型的计算方法研究

2016 年 10 月博士生论文开题报告

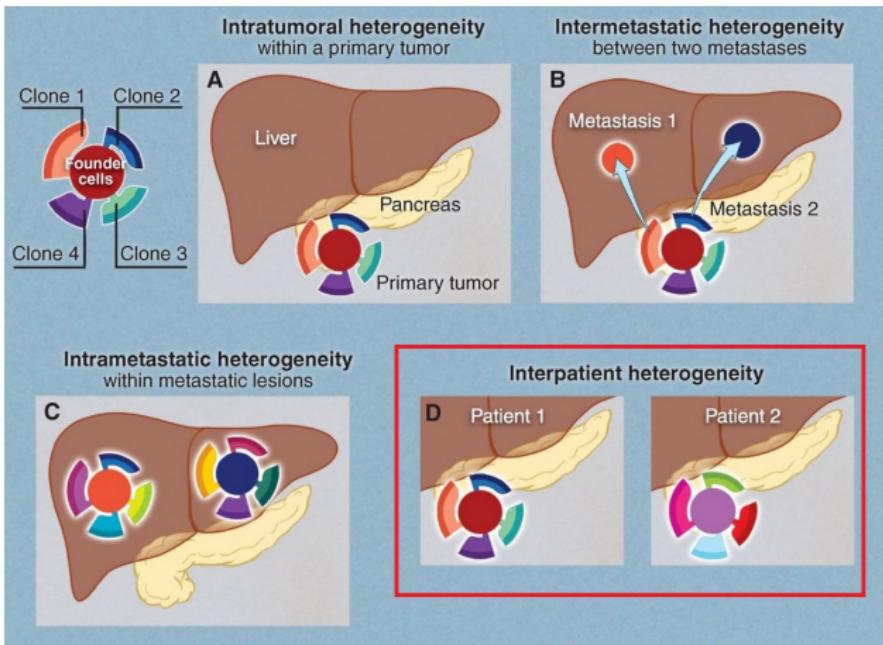
姓名: 王东方

学号: 2016310709

导师: 古槿



肿瘤异质性与肿瘤分类

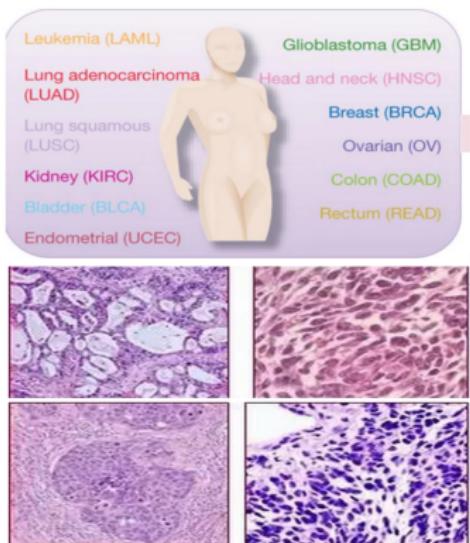


肿瘤分子异质性的四种类型 (Vogelstein et al., *Science*, 2013)



肿瘤分类:

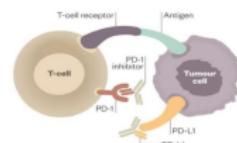
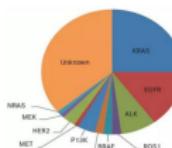
- 更好地理解肿瘤发生发展的机理;
- 针对不同类型设计特异性的治疗方案等



1.0 基于癌症发生部位

3.0 精准医学

分子特征：基因突变，免疫特性等



2.0 基于癌症病理特征

肿瘤分类方法的发展。



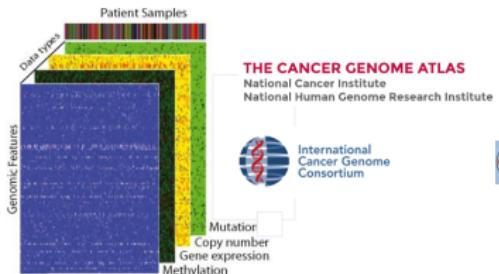
基于分子特征的肿瘤分类

- 对于肿瘤样本目前可获得从 DNA, RNA 到蛋白质的各个层次的分子信息；
- 不同层次的分子特征都包含与肿瘤分类相关的信息：
 - 不同的致癌突变 ([Ciriello et al., Nat. Genetics, 2013, etc.](#))
 - CpG 岛超甲基化定义了 CIMP(CpG island methylator phenotype) 子类型 ([Issa et al., Nat. Rev. Cancer, 2004, etc.](#))
 - ER/HER2/PR 定义乳腺癌的不同子类型 ([Cancer Genome Atlas, Nature, 2012, etc.](#))

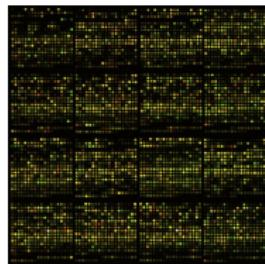


计算相关的问题

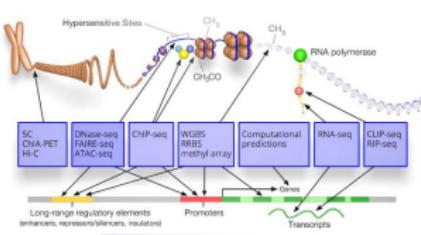
- 可获得的肿瘤相关数据集:



多组学肿瘤样本数据集: TCGA, ICGC (约 $10^2 \sim 10^3$)



肿瘤样本基因表达数据: GEO(约 $10^3 \sim 10^5$)



<https://www.encodeproject.org/>

肿瘤细胞系的更多组学数据: CCLE, ENCODE



- hubgene gene sets** are coherently expressed gene sets derived by aggregating many PMSIG350 hubgenes across individual biological states or processes.
- C1 positional gene sets** for each human chromosome and chromosome band.
- C2 enriched gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
- C3 model gene sets** based on conserved orthologs, mostly from a comparative analysis of the mouse, zebrafish, rat, and dog genomes.
- C4 coexpressional gene sets** defined by mining gene collections of co-expressed microarray data.
- C5 GO gene sets** consist of genes annotated to the same GO terms.
- C6 oncogenic signatures** defined directly from microarray gene expression data from cancer cell perturbations.
- C7 transcriptional signatures** defined directly from microarray gene expression data from immunologic studies.

生物知识数据集: COSMIC, MSigDB



计算相关的问题

- 从分子水平研究肿瘤分类的特点与困难：

- 影响因素复杂，高维度、多层次、多类型
- 采样不充分，样本量有限，多次采样样本异质性
- 可重复性较差



内容

① 整合多批次数据集的肿瘤分类方法

- 引言与现有方法综述
- 基于生物网络的“概念 - 观测”双层结构模型
- 下一步工作计划

② 整合多组学数据的肿瘤分类方法

③ 研究计划与工作安排



内容

① 整合多批次数据集的肿瘤分类方法

- 引言与现有方法综述
- 基于生物网络的“概念 - 观测”双层结构模型
- 下一步工作计划

② 整合多组学数据的肿瘤分类方法

③ 研究计划与工作安排



背景

- 基于组学的癌症研究在临床诊疗中取得进展有限；
- 很多方法与发现的**可重复性**较差。
 - 目前对某些癌症已经有很多数据集：
多批次、多中心的异质性与一致性；
 - 组学数据与**机理模型、先验信息**的整合。



现有方法综述

- 策略一：数据集拼接

将所有数据集组成一个组合矩阵

关键：“消除”批次影响 (batch effect)

(J Guinney et al., *Nature Medicine*, 2015; Johnson WE et al., *Biostatistics*, 2007; etc.)

- 策略二：聚类结果的再聚类

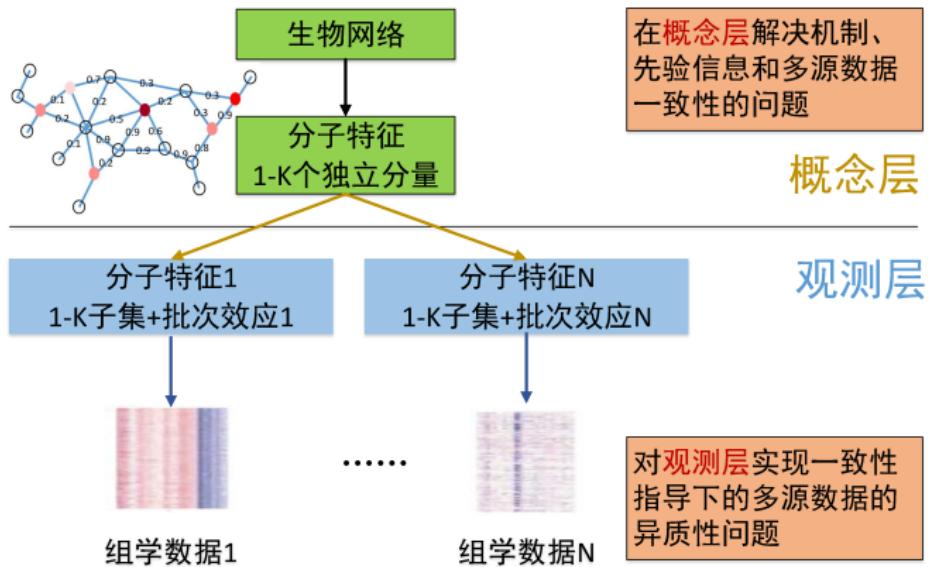
比较不同数据集得到的“类”的“相似性”

关键：相似性计算是基于数据集得到的信号特征
(signatures)

(Hoshida Y, et al. *Cancer Research*, 2009; CR Planey, et al., *Genome Medicine*, 2016.)



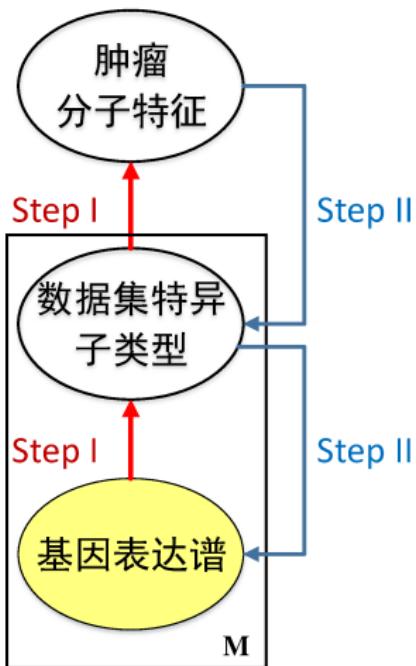
概念 - 观测双层结构



基于生物网络的“概念 - 观测” 双层结构



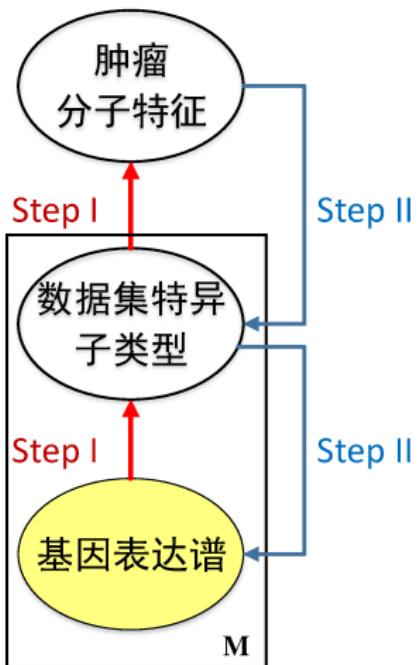
目前方法流程



- I. 基于“网络模块”的跨数据集的肿瘤分子特征:
- 单数据集聚类，得到各类的代表基因；
 - 结合文献挖掘基因、PPI 网络得到网络模块。



目前方法流程



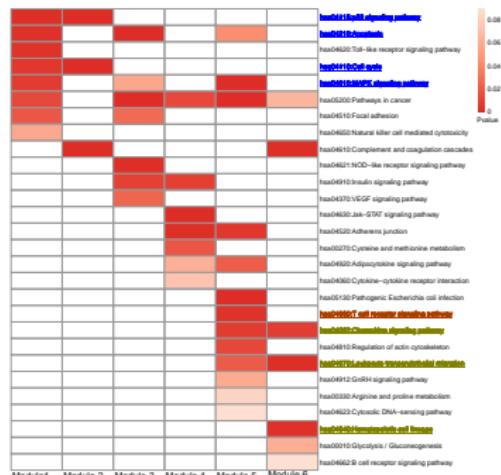
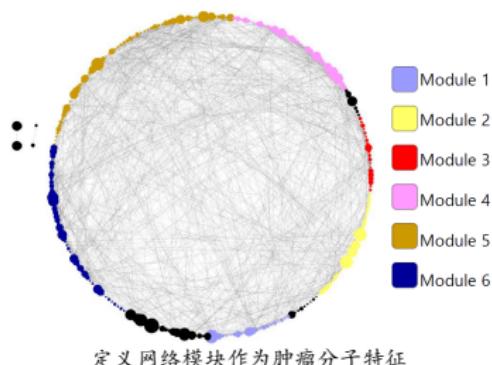
- I. 基于“网络模块”的跨数据集的肿瘤分子特征:
 - 单数据集聚类，得到各类的代表基因；
 - 结合文献挖掘基因、PPI 网络得到网络模块。
- II. 基于跨数据集的分子特征的分类:
 - 根据网络模块对单数据集的类中心重新定义；
 - 依据各数据集重新定义的类中心重分样本。



初步结果

数据: 12 个肝癌 (HCC) 基因表达数据集.

1. 网络模块及其功能注释:



使用 DAVID(Huang DW, et al., *Nature Protoc.*, 2009) 对信号模块进行功能注释



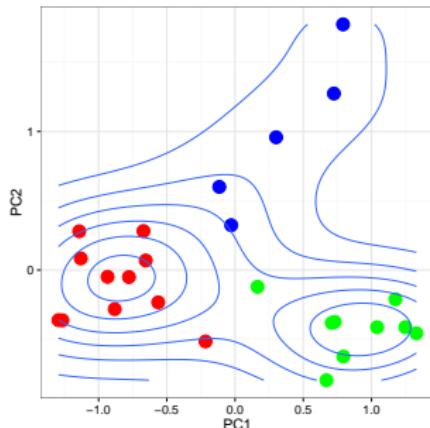
2. 基于网络模块的各数据集类中心定义:



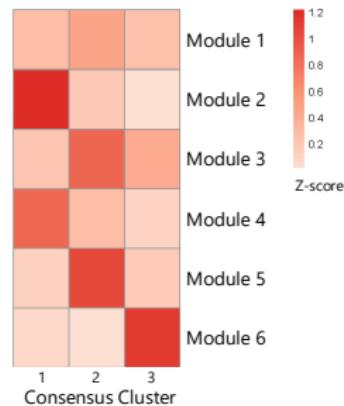
各数据集原分类中心在功能模块上的扰动程度。



3. 多数据集类中心的聚类:



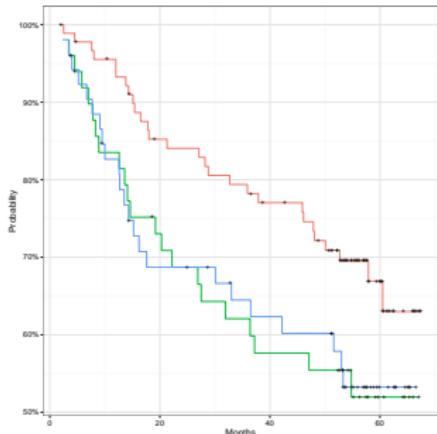
各个数据集各类扰动向量的主成分分析结果。



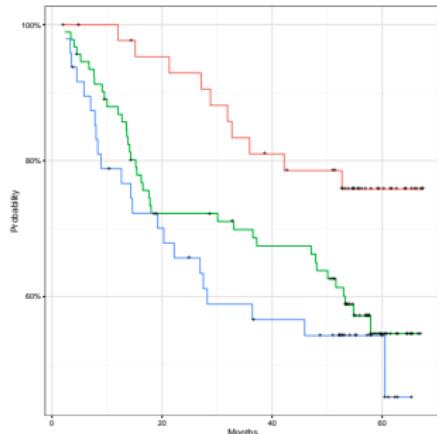
对扰动向量进行聚类，得到的类中心。



4. 样本的重新分类:



单数据集聚类的生存期分析结果
(HCCDB6). cox: 0.05



重新分类以后的生存期分析结果
(HCCDB6). cox: 0.008



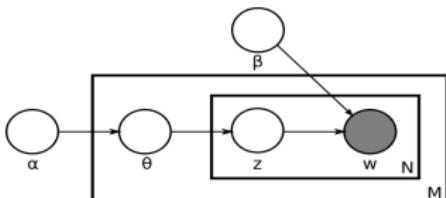
总结

- 提出一种基于多批次多中心的数据集对单一肿瘤进行分类的方法。
- 核心是基于生物网络的概念 - 观测双层结构。
 - 获得多数据集稳定存在的肿瘤分类相关分子特征；
 - 考虑多个数据集的异质性与一致性；
 - 尽可能的整合先验信息。
- 仍需要在多种癌症数据集上验证该模型，与现有方法比较。



下一步工作计划

1. 将现有模型发展成概率产生式模型：



(a) 经典 LDA 模型

COGNITIVE SCIENCE

Science 2015

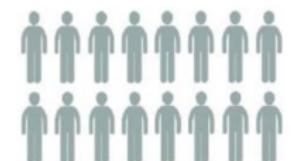
Human-level concept learning through probabilistic program induction

Brenden M. Lake,^{1*} Ruslan Salakhutdinov,² Joshua B. Tenenbaum³

(b) 概念学习模型



2. 解决小样本数据集的可靠分类问题：



小规模样本量新数据集的可靠分类



单样本的可靠分类?



数据库记录的样本及其分类





3. 结合多组学与细胞系数据：

- 基因之间的调控关系在肿瘤样本与细胞系之间是相对保守的，可以更好利用机制模型；(Seifert, et al., *Genome Biology*, 2016)
- 部分数据集在基因表达谱之外还有多组学数据，可以更好帮助肿瘤分类。



内容

① 整合多批次数据集的肿瘤分类方法

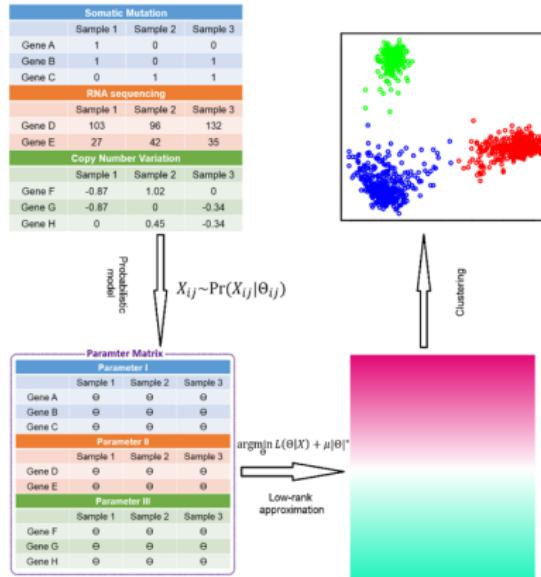
- 引言与现有方法综述
- 基于生物网络的“概念 - 观测”双层结构模型
- 下一步工作计划

② 整合多组学数据的肿瘤分类方法

③ 研究计划与工作安排



基于低秩近似的降维聚类 LRAcluster



$$\min_{\Theta} -\log \mathcal{P}(X|\Theta) + \mu |\Theta|_*$$

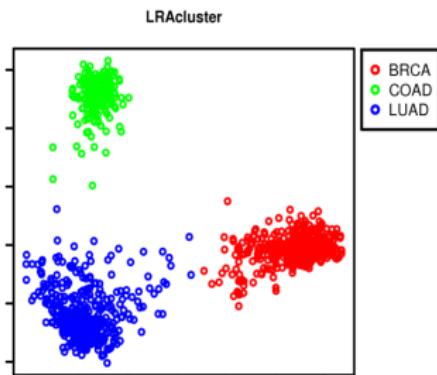
- X : 组数数据矩阵;
- Θ : 参数矩阵;
- \mathcal{P} : 概率分布: 高斯分布(实数), 二项分布(二值), 等。
- $||_*$: 核范数。

注: 与武丁明合作 (Wu et al., *BMC Genomics*, 2015), 参与设计算法和完成部分实验

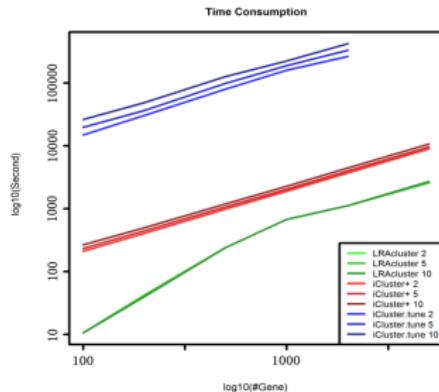


LRAcluster: 结果

1. 仿真: 三种不同癌症样本的混合



(c) LRAcluster 降至 2 维的结果



(d) LRAcluster 的时间性能

LRAcluster 仿真实验结果



2. 11 种癌症数据混合分析结果

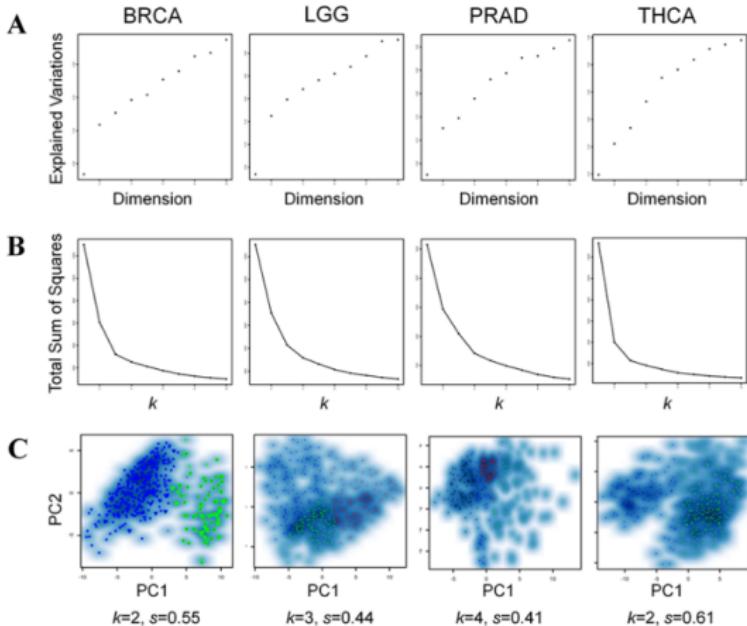
Table 1. The results of pan-cancer analysis.

φ	BRCA φ	COAD φ	GBM φ	HNSC φ	KIRC φ	LGG φ	LUAD φ	LUSC φ	PRAD φ	STAD φ	THCA φ	Total φ
C1 φ	1 φ	0 φ	0 φ	286 φ	0 φ	0 φ	0 φ	6 φ	0 φ	0 φ	0 φ	293 φ
C2 φ	0 φ	0 φ	0 φ	0 φ	0 φ	1 φ	0 φ	0 φ	0 φ	0 φ	411 φ	412 φ
C3 φ	0 φ	0 φ	41 φ	0 φ	0 φ	451 φ	0 φ	0 φ	0 φ	0 φ	0 φ	492 φ
C4 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	231 φ	0 φ	231 φ
C5 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	293 φ	0 φ	0 φ	293 φ
C6 φ	0 φ	190 φ	0 φ	1 φ	0 φ	0 φ	2 φ	0 φ	1 φ	0 φ	0 φ	194 φ
C7 φ	3 φ	17 φ	0 φ	0 φ	1 φ	0 φ	406 φ	7 φ	0 φ	0 φ	3 φ	437 φ
C8 φ	0 φ	0 φ	0 φ	0 φ	240 φ	0 φ	0 φ	0 φ	0 φ	0 φ	0 φ	240 φ
C9 φ	448 φ	0 φ	1 φ	2 φ	1 φ	0 φ	4 φ	1 φ	0 φ	0 φ	0 φ	457 φ
C10 φ	8 φ	1 φ	0 φ	195 φ	0 φ	0 φ	6 φ	60 φ	0 φ	0 φ	0 φ	270 φ
Total φ	460 φ	208 φ	42 φ	484 φ	242 φ	452 φ	418 φ	74 φ	294 φ	231 φ	414 φ	3319 φ

将 LRAcluster 应用到 pan-cancer 分析



3. 单种癌症数据混合分析结果



将 LRAcluster 应用到癌症分类分析



总结

- 提出一种基于低秩近似的分析超高维多组学数据的方法；
- 算法时间性能优良，能快速进行分析与降维可视化；
- 已经应用于 pan-cancer 与单种癌症数据集的分析。



内容

① 整合多批次数据集的肿瘤分类方法

- 引言与现有方法综述
- 基于生物网络的“概念 - 观测”双层结构模型
- 下一步工作计划

② 整合多组学数据的肿瘤分类方法

③ 研究计划与工作安排



计划与时间安排

- 2016.8 - 2017.1: 完成多批次数据整合现有模型的相关整理工作；
- 2017.1 - 2017.8: 完成基于概率产生式模型的多批次数据聚类工作；
- 2017.8 - 2018.7: 完成多批次数据与癌症细胞系和多组学数据整合的相关工作；
- 2018.7 - 2019.1: 完善已有工作，开发整合分析平台；
- 2019.1 - 2019.4: 撰写博士生论文。



已发表文章

- Wang, D., Gu, J., Wang, T., & Ding, Z. (2014). OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics*, 30(15), 2237-2238.
- Wu, D., Wang, D., Zhang, M. Q., & Gu, J. (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC genomics*, 16(1), 1.
- Wang, D., & Gu, J. (2016). Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1), 58-67.



谢谢!

请各位老师批评指正

