深度学习和GSMM

https://doi.org/10.1371/journal.pcbi.1007084

1. 摘要

1️⃣一些相对独立发展的领域

1. 组学数据在爆炸增长
2. 统计学和机器学习是解释生物表型(phenotypes)的主要计算方法
3. 基于约束的代谢模型是解释基因型-表现型-环境的主要方法

然而这些领域在生物信息医药方向的融合却鲜有研究

2️⃣本文要干什么：

1. 描述：如何融合机器学习和限制型GSMM
2. 综述：上述两个领域的最新进展&数理和实践
3. 最后：提出一种新的融合框架

2. Introduction

2.1. 背景1：组学数据的指数增长

1️⃣高通量技术(high-throughput technologies)的发展促进了数据的收集，尤其是组学(Omic)数据

2️⃣主要的组学和应用

技术领域	技术手段	中文翻译
Genomics	DNA sequencing	基因组学：DNA测序
Transcriptomics	Microarrays and RNA sequencing	转录组学：微阵列和RNA测序
Epigenomics	DNA methylation and histone modifications	表观基因组学：DNA甲基化和组蛋白修饰
Proteomics	Protein mass spectrometry	蛋白质组学：蛋白质质谱分析
Metabolomics	Metabolite mass spectrometry	代谢组学：代谢物质谱分析

3️⃣通过组学数据可以直观分析基因改变和细胞活动

4️⃣大量的组学数据急需适当的分析

2.2. 背景2：机器学习和数学模型

2.2.1. 机器/深度学习用于处理组学数据

1. 机器学习：输入→学习归纳输入→通过经验提高预测准确性
2. 机器学习在计算/生物信息的优势
    1. 预测能力强
    2. 机器学习所需的假设更少(比如PINN)
3. 应用举例：理解RNA折叠，预测突变对剪切的影响，研究基因表达谱

2.2.2. 生物分子的数学模型

1. 组学领域分析其实分为数据驱动，还有假设驱动
2. 假设驱动方法是很拉跨的，在于难以确定潜在的生物机制
3. 这一点限制有一个例外，就是基于约束的代谢模型(CBM)

2.2.3. 二者模型的结合

1. 上述两个模型往往都被单独使用
2. 二者计算特性的相似使得完全可以结合

2.3. 文章核心思想

1. GSMM可以用来生成通量组学(fluxomic)数据，也就是额外的omic层，并于现有组学数据整合(但这不是最佳方案)
2. 分为两种模型
    1. 数据驱动：机器学习和深度学习
    2. 知识驱动：CBM(基于限制的GSMM)
3. 相比单纯用机器学习，引入GSMM和组学数据更牛逼

3. 数据驱动的生物分子模型

一言以蔽之：机器学习可以在特定的环境和组学类型，从大量的组学数据中，提取有用的知识

3.1. 机器学习的方法

1. 监督式：预测给定样本的目标，如ANN/SVM
    1. 分类器：预测样本类别，如致病性和非致病性
    2. 回归器：估计数字的量，如致病性水平
2. 非监督式：变量/相关性分解
    1. 关联算法：发现样本中潜在的规则和趋势
    2. 聚类算法：更具样本固有的隐形特征来对样本分类，如K-均值聚类和分层聚类
3. 组学数据的压缩/简化预处理

1. 主成分分析PCA：将数据降维为低维表示，总结变量间的最大方差
2. 因子分析，根据变量间相关性分解数据
3. 矩阵因式分解，将数据矩阵分解为去噪成分

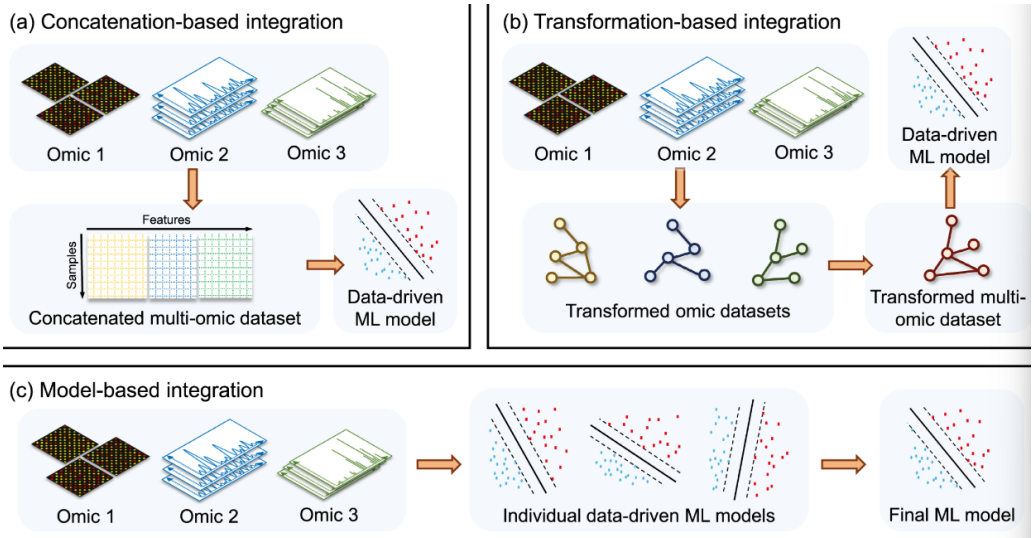
3.2. 多组学数据与机器学习

1️⃣数据整合方法：

1. 可促进多个组学数据集的分析，更密切反应表现型-基因型关系
2. 不同组学之间相互关联，单个数据缺失可从多组数据中得到补偿

2️⃣大规模数据整合：元维度方法(metadimensional methods)

1. 特点：跨越多个数据源，能应对变量输入
2. 分类：基于连接/转换/模型的整合respectively早期/中期/晚期模型



3. PS：机器学习中，处理跨越多个数据源的算法aka多模态/多视角学习算法

3️⃣基于连接的集成

1. 原理：将所有组学数据合并成一个大矩阵(多种不同的数据硬揉在一起)，在数据机器学习模型
2. 缺点：每种数据都有故有偏差
3. 归一化(Normalization)技术：让不同数量级数据趋于一个比例，但噪声和方差仍然会影响

4️⃣基于转换的集成

1. 原理：将不同组学数据转化为统一数据格式(中间形式)，合并成一个融合数据集，再输入机器学习模型
2. 特点：保留了数据的原始特性，但是忽略了数据间交互特性

5️⃣基于模型的集成

1. 原理：每个组学数据先单独用机器学习学习，合并学习得到的数据
2. 特点：对于拟合敏感，只适用于数据池极其异构的情况

4. 基于限制的生物分子网络分析

4.1. GSMM(Genome-scale metabolic models)

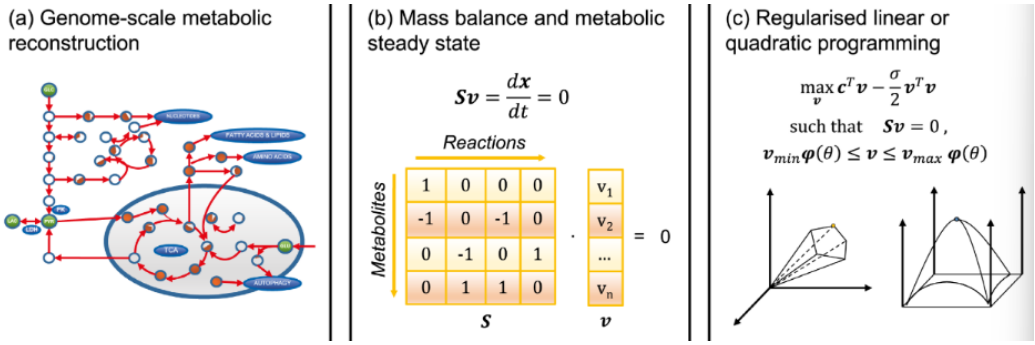
4.1.1. 概述

1. 在代谢组分析不可行的情况下，它们也能评估细胞群的完整代谢状态
2. GSMM的两个基本假设
    1. 质量守恒：输出给外部的产物质量，等于输入内部的消耗底物的质量
    2. 拟稳态：内部各代谢物质量动态平衡
3. GSMM和CBM
    1. GSMM：只适用于稳态条件，但是可以涵盖整个细胞代谢
    2. CBM：计算成本高，可以动态描绘小型代谢系统

4.1.2. 代谢通量的建模

1. FBA通量平衡分析，不必多说
2. 目的：确定目标反应物的最大/最小通量配置，当然反应量通量也可以换成生物量通量
3. 形式：针对反应通量子集的线性优化问题

4.1.3. 基于约束的数据整合和通量组生成



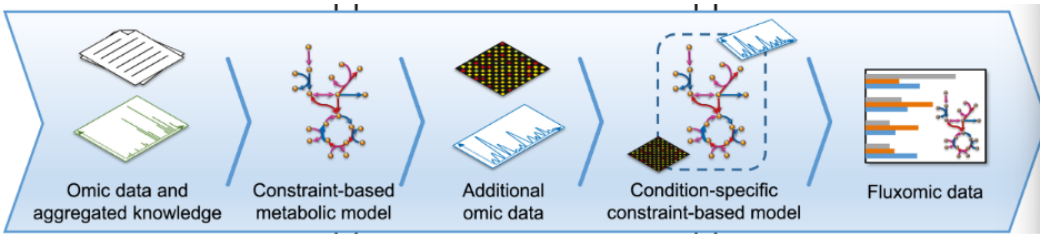
1. 构建一个人工编辑的GSMM，用于记录细胞中发生的所有反应
2. 构建化学计量矩阵S，每种物质的通量向量v，拟稳态的限制即可表达为$S \cdot v = 0$
3. 正则优化(regularized)：目标函数中减去一个凹函数(凸优化？？？)

4.2. 基于特定条件的约束模型

4.2.1. 背景

1. GSMM 中，反应的数量通常多于代谢物的数量，若无限制则问题的确定性太低(解空间太大)
2. 增加物理/生物/化学约束，可使模型更接近生物学意义
3. 限制诸如：酶活性，代谢物耦合，转录调控等

4.2.2. 特定环境(限制)GSMM的构建



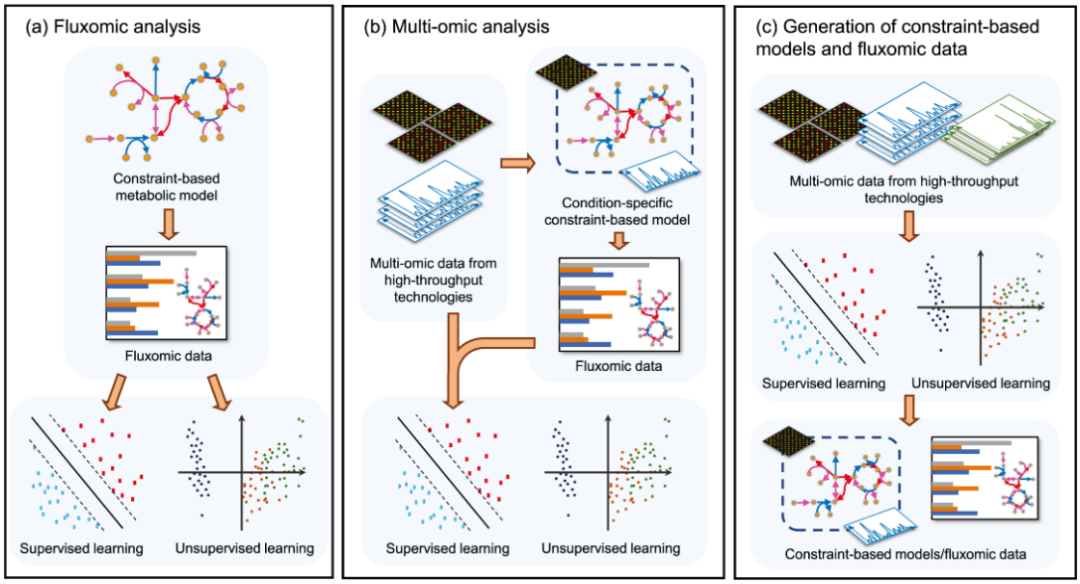
4.2.3. 组学数据的整合(上图第三步)

1. 转录图谱的整合(Transcriptional profiles)：修剪代谢网络的算法
    1. 基于开关的方法：为基于表达设定阈值，低于阈值的基因表达就关闭，从GSMM中剪去
    2. 基于阀门的方法：以连续的方式将转录信息映射到基于约束的模型上
2. 基因和蛋白质表达数据的整合：较少，就不一一列举了

5. 限制模型-机器学习

5.0. 概览(潜在优势)

1. 遗传和环境扰动以非线性方式在代谢网络中传播，最终体现在代谢通量
2. GSMM分析获得的通量可被视为另一个组学层，可通过学习算法分析
3. 将二者结合的研究鲜有，但是2019年以前的大致有这些，可分类为



1. 有监督/无监督通量组学分析：在通用GSMM上执行FBA或相关技术→获得通量数据→用作无监督或有监督机器学习的输入
2. 有监督/无监督多组学分析：用高通量分析技术获得多组学数据集→不同源数据归一化→用作无监督或有监督机器学习的输入
3. 基于约束的模型生成和通量组学数据：直接把组学数据丢进去，生成或者改进GSMM

5.1. 有监督的通量组学分析

1️⃣概述：通过通用GSMM获得预测的生物目标→获得FBA等技术的输出→将输出直接输入机器学习进行有监督分析

2️⃣研究1：从细胞内通量配置推断细菌的生长条件

1. 使用内部代谢通量作为输入，对特定 FBA 解决方案的生长条件进行预测
2. 正则化选择最相关的通量，防止过度拟合

3️⃣研究2：正确识别抑制性药物的副作用

1. 敲除硅学有关基因，重构GSMM，以此模拟药物特性
2. 通过通量变动分析(FVA)估算代谢通量的扰动
3. 获得结果输入SVM

3️⃣研究3：深度神经网络和差分搜索算法被用于设计大肠杆菌基因缺失干预措施

4️⃣研究4：预测不同生物处理设置下的滴度、生产率和产量

5️⃣研究5：基于代谢网络约束的判别分析技术--动态基本模式回归判别分析，即最能区分实验条件的通路激活模式

5.2. 无监督的通量组学分析

1️⃣用途：

1. 未明确生物目标时，可用无监督学习描述多个样本之间的差异性和相关性，从而对新陈代谢聚类
2. 采用降维技术来解构与基于约束的模型相关的整个通量空间，捕捉通路间微不足道的交叉相关性
3. 利用化学计量学约束学习目标

2️⃣研究1：GSMM探索酵母新陈代谢中的外显性

1. 定义适合度：FBA生长速率比
2. 对所有参与新陈代谢的基因的单倍和双倍有害突变体的适合度景观进行聚类

5.3. 有监督的多组学分析

1️⃣概述：

1. 多组学特征集：结合CBM生成的通量组学数据和其他多组学数据
2. 通过机器学习预测感兴趣的目标

2️⃣研究：大肠杆菌的代谢本质

1. FBA方法+人工敲除基因，可以有效预测重要反应
2. 反应重要性的评估往往只基于生物量积累率
3. 改进1：使用SVM分类器，结合其他组学数据，提高FBA预测
4. 改进2：使用FBA适配性和遗传相互作用得分，训练随机森林

3️⃣建立特定条件下的代谢模型来反映细胞不同环境下的代谢能力：如利用线性编程模型来强制基因表达-代谢通量一致，再用KNN对模型二元分类

4️⃣ 充分利用多种组学分析方法：

1. 创建特定条件的 GSMM
2. 基于机器学习的数据整合来实现两阶段整合

5️⃣ 更复杂的数据整合管道：如DeepMetabolism(ANN方法)，将无监督预训练与有监督训练相结合，建立一个具有预测表型结果能力的深度学习模型

具体看文献吧，不一一总结了

5.4. 无监督的多组学分析

1️⃣ 用途1：可应用于由实验数据和GSMM生成的异质数据学

PS：异质数据集——不同类型或来源的数据组成的数据集

2️⃣ 用途2：将GSMM和代谢作为理解基因组变异的基础，例如GEESE(基因表达潜在空间编码器)

5.5. 生成基于约束条件的模型和通量组数据

1️⃣ 概述：

1. 除了分析通过 CBM 生成的通量组之外，还可以将机器学习与 CBM 本身相结合，以获取新的通量组信息
2. 例如用SVM-kNN-决策树预测通量分布

这都是些啥。。。。先这样吧