生物信息学中的数学方法

Jingbo Xia

Huazhong Agricultural University

xiajingbo.math@gmail.com

2023-12-07

Table of contents I

Outline

发展中的生物信息学和其采用的数学方法 I

参考资料:

1. "What should I study if I want to learn bioinformatics?" Quora, https://www.quora.com/How-much-math-is-required-for-bioinformatics
2. 《生物信息学及其主要数学算法》吴春艳，王靖飞
3. 《An Introduction to Bioinformatics Algorithms》Neil C. Jones and Pavel A. Pevzner
4. 《生物信息学与系统生物学》中科院数学与系统科学研究院。http://doc.aporc.org/wiki/Course001
5. "如何自学生物信息学 "知乎 https://www.zhihu.com/question/20543692

发展中的生物信息学和其采用的数学方法 II

Many people believe that you need to be good at math to study bioinformatics. However, this is not necessarily true. There are many bioinformatics tools and resources that do not require a strong math background. For example, there are many web-based tools that allow you to visualize and analyze biological data. [a]

[a] Quora, 2018 年 3 月 15 日

发展中的生物信息学和其采用的数学方法 III

1954 年 Crick 提出了遗传信息传递的规律，DNA 是合成 RNA 的模板，RNA 又是合成蛋白质的模板，称之为中心法则 (Central dogma)，这一中心法则对以后分子生物学和生物信息学的发展都起到了极其重要的指导作用。
1956 年美国田纳西州盖特林堡召开的“生物学中的信息理论研讨会”，首次产生了生物信息学的概念。
1963 年 Nirenberg 和 Matthai 通过实验研究，编码 20 氨基酸的遗传密码得到了破译。限制性内切酶的发现和重组 DNA 的克隆（clone）奠定了基因工程的技术基础。正是由于分子生物学的研究对生命科学的发展有巨大的推动作用，生物信息学的出现也就成了一种必然。
20 世纪 80 年代末随着人类基因组计划的启动而兴起一门新兴学科——基因组信息学，后改为生物信息学。1987 年林华安博士正是称这一领域为“生物信息学（Bioinformatics)”。近年来，计算机和因特网的快速发展更是为生物信息的传递　供了硬件基础和便利条件。(生物信息学的实质就是运用计算机科学及网络技术来解决生物学问题。)
2001 年 2 月，人类基因组工程测序的完成，使生物信息学走向一个高潮。

发展中的生物信息学和其采用的数学方法 IV

Subject	4 Exhaustive Search	5 Greedy Algorithms	6 Dynamic Programming Algorithms	7 Divide-and-Conquer Algorithms	8 Graph Algorithms	9 Combinatorial Pattern Matching	10 Clustering and Trees	11 Hidden Markov Models	12 Randomized Algorithms
Mapping DNA	o								
Sequencing DNA					o				
Comparing Sequences			o	o		o			
Predicting Genes			o						
Finding Signals	o	o						o	o
Identifying Proteins					o				
Repeat Analysis						o			
DNA Arrays					o				
Genome Rearrangements		o							
Molecular Evolution							o		

a

[a] 《An Introduction to Bioinformatics Algorithms》 Neil C. Jones and Pavel A. Pevzner

发展中的生物信息学和其采用的数学方法 V

与生物信息学相关的数学方法

1. 概率论与随机过程理论，如隐马尔科夫链模型（Hidden Markov Model, HMM）：数据库的搜索、序列比较、建立蛋白质模型及发现新基因
2. 统计学，包括多元统计学，是生物信息学的数学基础之一.
3. 运筹学，如动态规划（Dynamic Programming）：序列比对。
4. 信息论：在分子进化、蛋白质结构预测、序列比对中有重要应用
5. 拓扑学，尤其是几何拓扑：DNA 超螺旋研究，多肽链折叠研究；
6. 函数论（如傅里叶变换和小波变换）：生物信息学中的常规工具；
7. 计算数学 (如常微分方程数值解法): 分子动力学基本工具.
8. 组合数学: 分子进化和基因组序列研究.
9. 群论：研究遗传密码和 DNA 序列的对称性.
10. 人工神经网络方法：核酸和蛋白质序列的分析。例如，转录终端预测，启动子、外显子和内含子的鉴别；转录控制信号分析，DNA 曲率分析。
11. 最优化理论与算法: 蛋白质空间结构预测和分子对接研究.

发展中的生物信息学和其采用的数学方法 VI

生物信息学:

生物信息学-1. 基础知识
生物信息学-2. 序列分析与比对
生物信息学-3. 单体型组装与推断
生物信息学-4. 蛋白质结构预测
生物信息学-5. 蛋白质结构比较
生物信息学-6. 功能预测与注释
生物信息学-7. 基因组变异分析
生物信息学-8. 高通量技术
生物信息学-9. 新一代测序技术

计算系统生物学

计算系统生物学 1. 概述
计算系统生物学 2. 基因调控网络推断
计算系统生物学 3. 转录调控网络
计算系统生物学 4. 转录因子合作网络
计算系统生物学 5. 网络分析;
计算系统生物学 6. 网络比对
计算系统生物学 7. 活性通路

发展中的生物信息学和其采用的数学方法 VII

生物信息学:

生物信息学-1. 基础知识
生物信息学-2. 序列分析与比对
生物信息学-3. 单体型组装与推断
生物信息学-4. 蛋白质结构预测
生物信息学-5. 蛋白质结构比较
生物信息学-6. 功能预测与注释
生物信息学-7. 基因组变异分析
生物信息学-8. 高通量技术
生物信息学-9. 新一代测序技术

计算系统生物学

计算系统生物学 1. 概述
计算系统生物学 2. 基因调控网络推断
计算系统生物学 3. 转录调控网络
计算系统生物学 4. 转录因子合作网络
计算系统生物学 5. 网络分析;
计算系统生物学 6. 网络比对
计算系统生物学 7. 活性通路

生物信息学与计算系统生物学 (Bioinformatics and Computational Systems Biology)
时间: 2012 年 2 月 22 日-2011 年 4 月 27 日, 每周三、五下午 13:30-16:10
地点: 中科院研究生院中关村教学楼 N215 教室

发展中的生物信息学和其采用的数学方法 VIII

[From Wikipedia] Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.

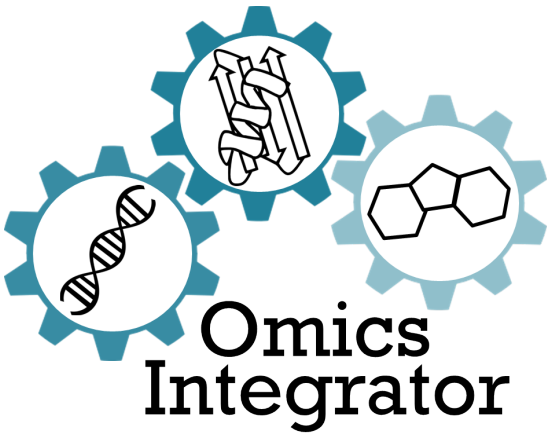
发展中的生物信息学和其采用的数学方法 IX

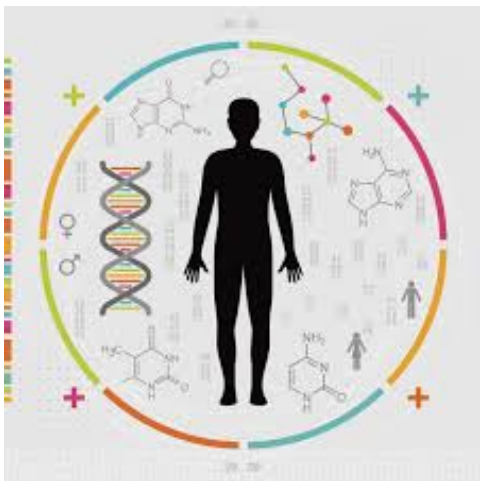
Contents [hide]

发展中的生物信息学和其采用的数学方法 X

大数据和人工智能新时代下的一些趋势

1. 多组学 (Multi-Omics) 数据的蓬勃发展——契合大数据时代
2. 数据的个性化和精准化——生物、医药等关键突破
3. 生物对象的类别和量的刻画——统计学，以及深度学习等革命性方法
4. 生物对象两两之间的关联分析，和因果推断——数据挖掘
5. 系统观点下的生物知识网络——知识图谱



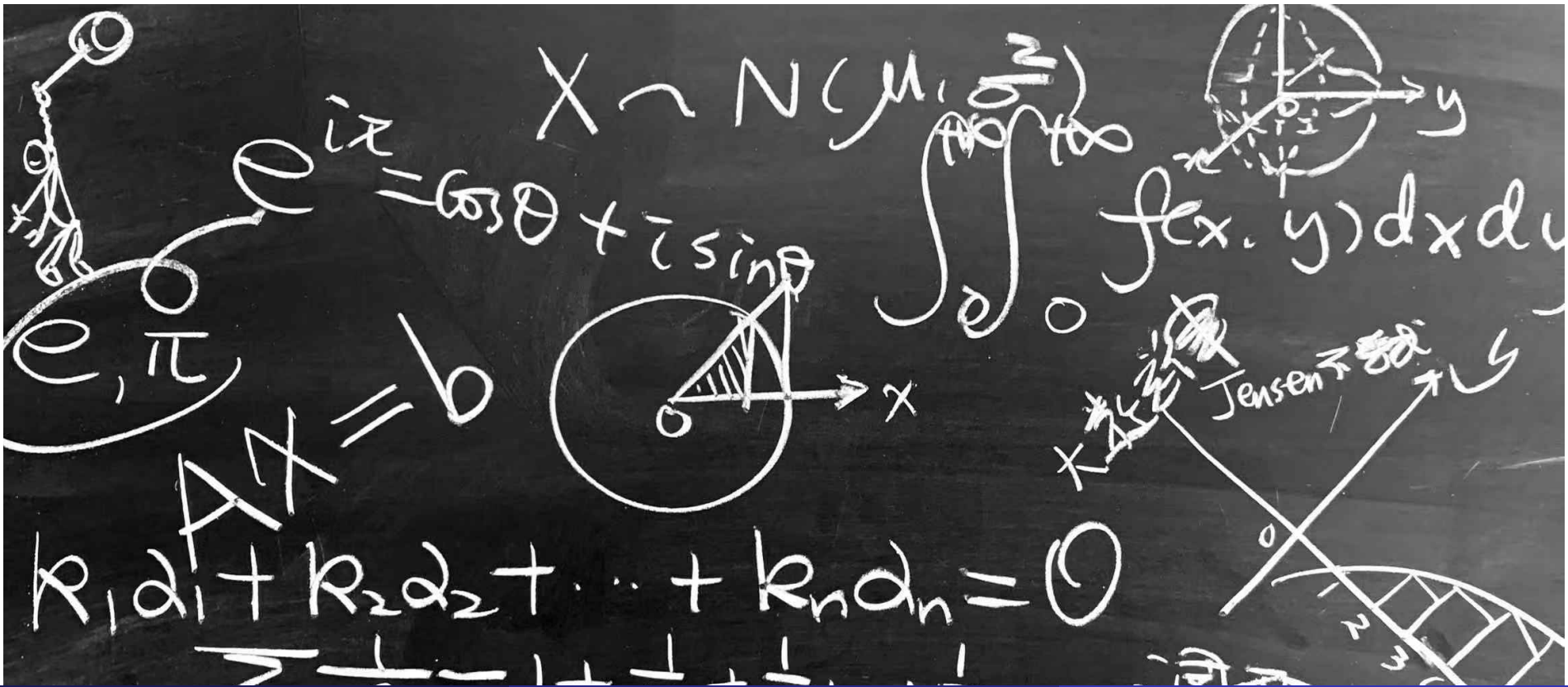


Outline

数学理论和方法的学习

大纲中的数学课程

1. 《微积分》连续的量
2. 《线性代数》离散的量
3. 《概率论和统计学》随机的量
4. 《生物信息数学基础》
5. 《离散数学》，《多元统计》，《数据挖掘》，《生物文本挖掘与知识发现》。。。



数学理论和方法的学习

大纲中的所有课程

1. 生物类。。。
2. 化学类。。。
3. 计算机类。。。
4. 数学类。。。
5. 。。。

数学理论和方法的学习

有关数学理论和方法的学习，不二法则

1. 打好数学基础
2. 交叉融合，“学 “以致 “用”。

Outline

一个例子《Hyper Geometric Test》

抽球问题和超几何检验。
在富集分析中的应用，包括 GO 富集分析，通路富集分析，等等。

一个例子《Hyper Geometric Test》 I

Background data.
Assume m out of the whole N backgroud genes are relevant with a targeted pathway.

A wet-lab or a dry-lab test
During a lab test, e.g., GWAS, k genes are pinpointed. Among k genes, x of them are relevant with the mentioned pathway.

Prob question.
How to compute the probability of this instance?

请思考。

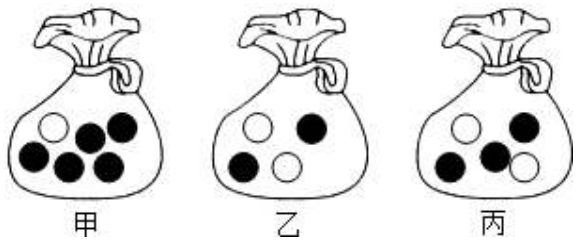
一个例子《Hyper Geometric Test》 II

Answer sheet:

$$P(x) = \frac{C_m^x \times C_{N-m}^{k-x}}{C_N^k}. \tag{1}$$

注[1]。

[1]事实上，这个公式和抽球问题相同。在一个袋子中有 N个球，其中黑球 m个。现在，在一次实验中，抽出了 k个球，其中有 x个黑球。



一个例子《Hyper Geometric Test》 I

Enrichment analysis.
是否这次实验中获得的基因显著地富集在给定通路中？

请思考：选用何种数学模型来表征这个问题？

一个例子《Hyper Geometric Test》 II

使用假设检验。
原假设 $H0$是：_____。
备择假设 $H1$是：_____。

一个例子《Hyper Geometric Test》 III

使用假设检验。
原假设 $H0$是:
与随机抽取群体相比，所获基因中包含的通路相关基因在数量上无显著差异。
备择假设 $H1$是:
与随机抽取群体相比，所获基因中包含的通路相关基因在数量上存在显著差异。

p-值为 [2],

$$p = 1 - \sum_{i=0}^{x-1} P(i) = P(x) + P(x+1) + \cdots + P(k), \tag{2}$$

where $P(x)$ refers to eqn. (1).

[2] *clusterProfiler*, GO 富集分析包。Yu, Guangchuang. "Using clusterProfiler to identify and compare functional profiles of gene lists." (2013). School of Biological Sciences, The University of Hong Kong

一个例子《Hyper Geometric Test》

GWAS 基因筛选和通路富集

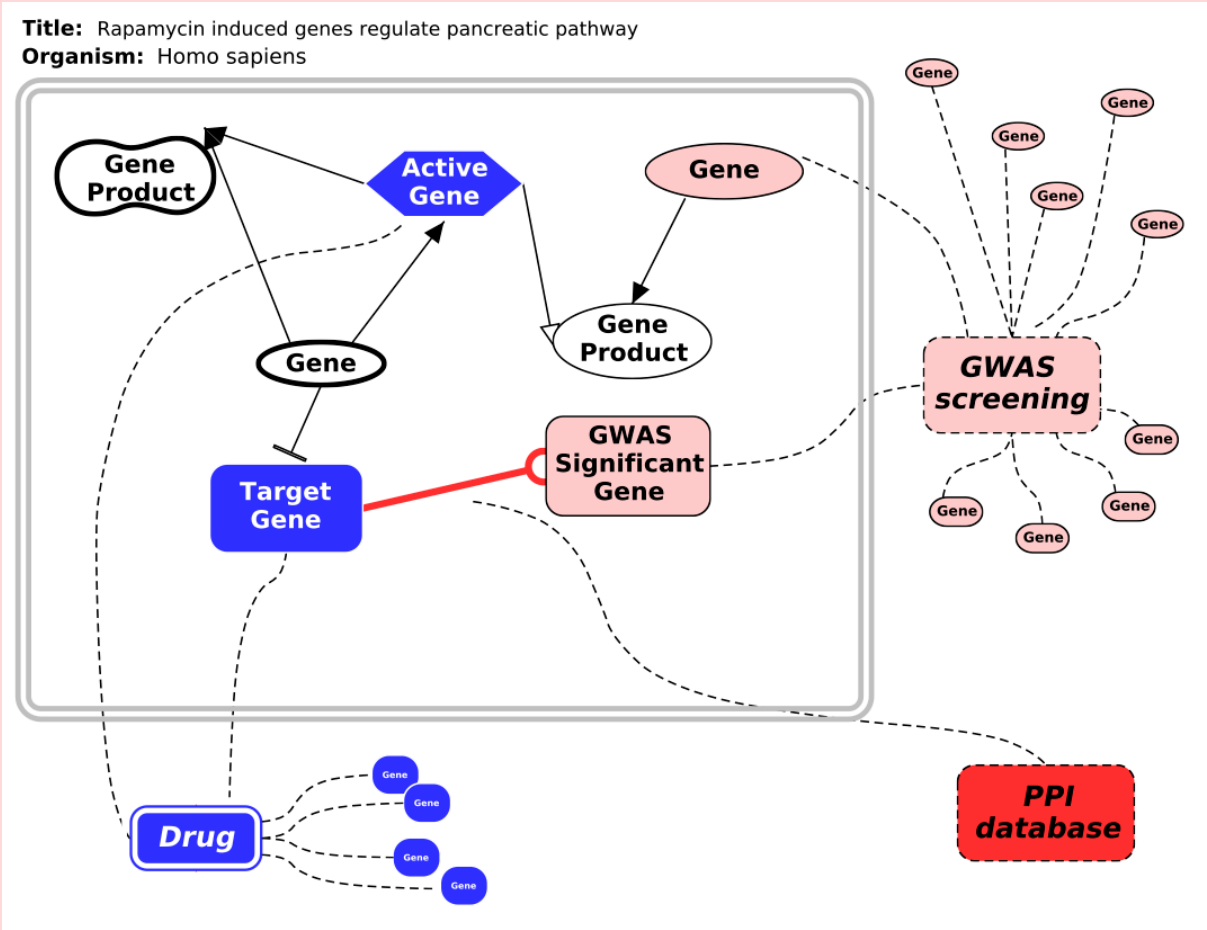
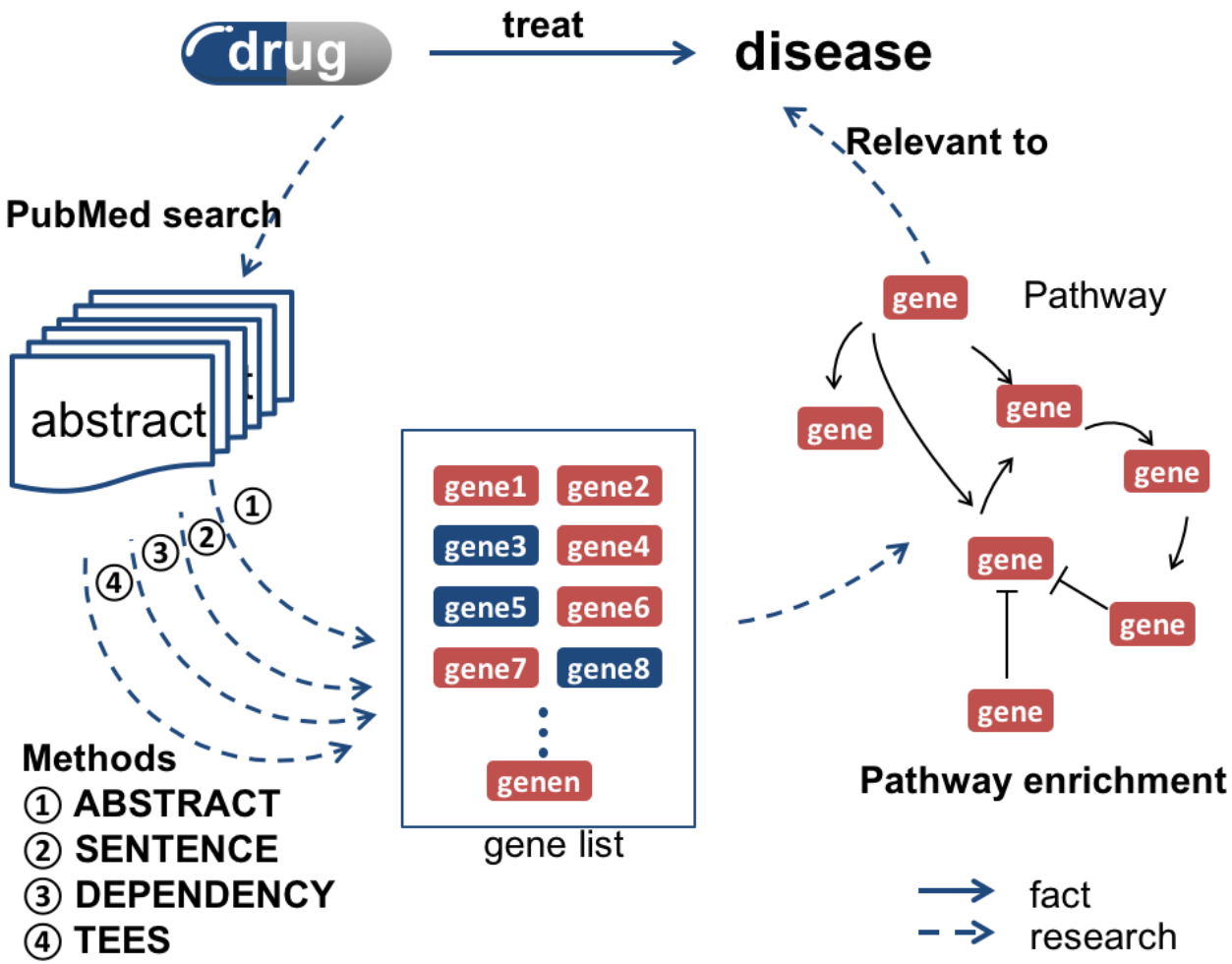


图 1: 通路富集示例: Rapamycin and pancreatic genes in pathway enrichment.

一个例子《Hyper Geometric Test》



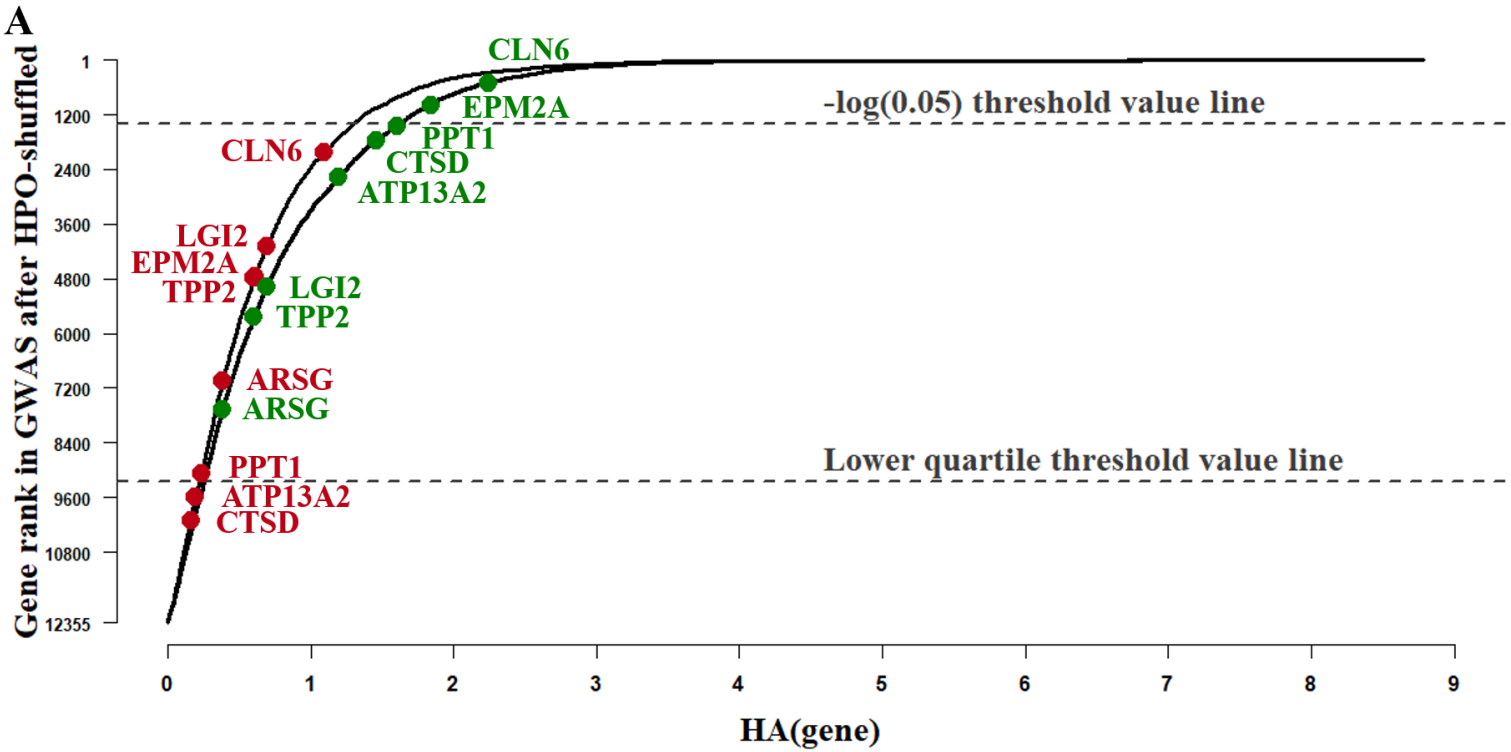
Some update [3] .

[3] Qin, Xuan, Xinzhi Yao, and Jingbo Xia. "A Novel Metric to Quantify the Effect of Pathway Enrichment Evaluation With Respect to Biomedical Text-Mined Terms: Development and Feasibility Study." JMIR Medical Informatics 9.6 (2021): e28247.

一个例子《The Wilcoxon Rank-Sum Test》

一个例子《The Wilcoxon Rank-Sum Test》

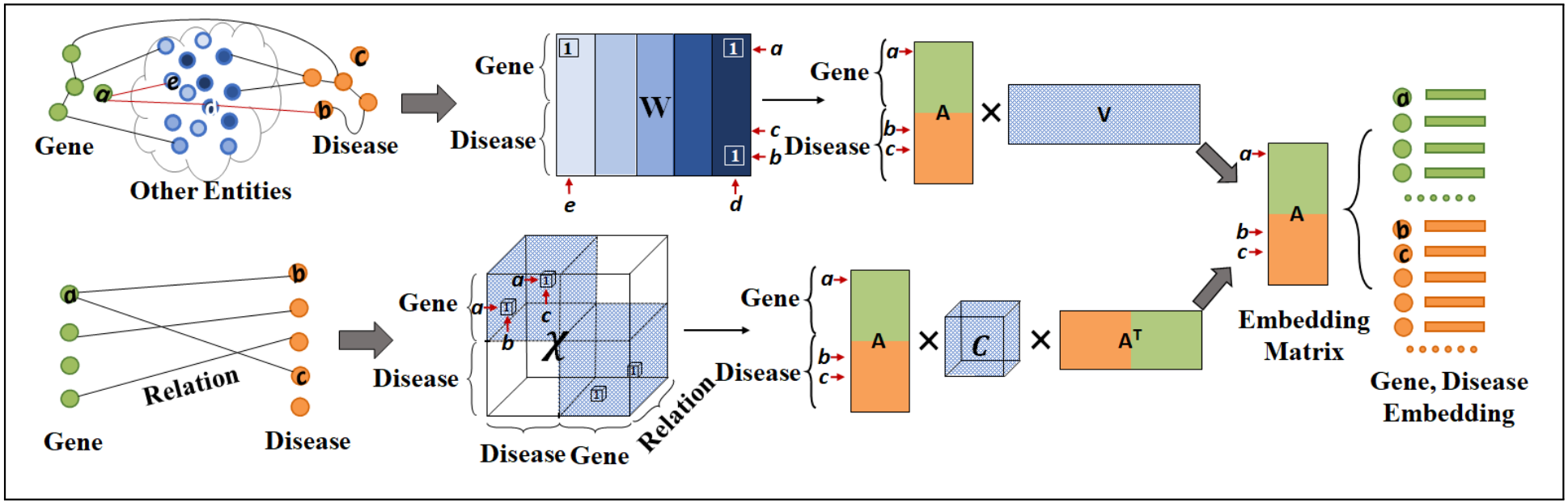
一个例子《HPO-Shuffle [4] 》
算法：Wilxocon 等秩和检验
应用场景：Gene prioritization.



[4]Wang, Shuguang, et al. "HPO-Shuffle: an associated gene prioritization strategy and its application in drug repurposing for the treatment of canine epilepsy." Bioscience reports 39, no. 9 (2019).https://portlandpress.com/bioscirep/article/39/9/BSR20191247/220424

一个例子《Data Fusion》

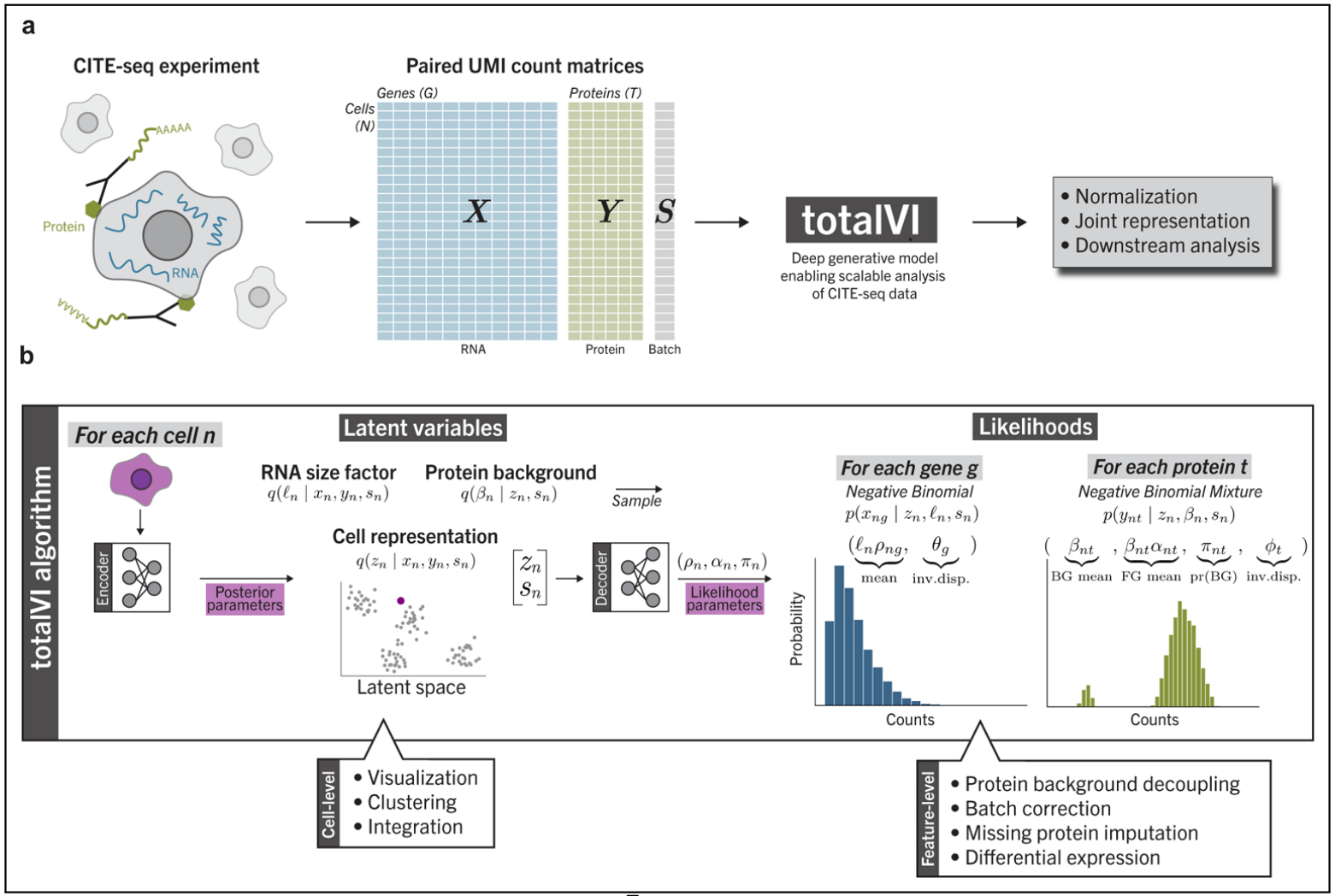
For G different genes and D different diseases under consideration, we assume there are two heterogeneous graph data in the form of uni-relation and multi-relation triples.



An algorithm to

- generate new core tensors with good approximation, by using heterogeneous matrix/tensor data.
- compute gene/disease embedding, from uni-relation and multi-relation triples.

一个例子《Data Reconstruction》



Data reconstruction of single-cell data [5].

[5] Gayoso, Adam, et al. "Joint probabilistic modeling of single-cell multi-omic data with totalVI." Nature Methods 18.3 (2021): 272-282. https://www.nature.com/articles/s41592-020-01050-x.pdf


The End