• 医学信息研究 •

统计学方法在生物信息学分析中的应用

刘 壮 张 悦

（中国医科大学《中国卫生统计》杂志 沈阳 110122）

〔摘要〕 介绍序列相似性分析、基因表达分析、基因转录调控网络分析和序列结构与模式识别分析中统计学方法的应用情况，为科研人员系统学习生物分析技术提供理论依据。

〔关键词〕 统计方法；生物信息；基因表达；序列结构

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.06.004

Application of Statistical Methods in Bioinformatics Analysis *LIU Zhuang, ZHANG Yue, Editorial Office of Chinese Journal of Health Statistics, China Medical University, Shenyang 110122, China*

〔**Abstract**〕 The paper introduces the application of statistical methods in sequence similarity analysis, gene expression analysis, gene transcriptional regulatory network analysis, sequence structure and pattern recognition analysis, providing theoretical basis for researchers to study systematic biological analysis techniques.

〔**Keywords**〕 statistical methods; bioinformatics; gene expression; sequence structure

1 引言

20 世纪 80 年代末人类基因组计划启动，基因组学测序数据迅猛增加，随之兴起生物信息学这门新的交叉学科。伴随生物学和医学的迅速发展，特别是人类基因组计划的顺利推进，产生海量生物学数据，特别是生物分子数据积累速度在不断快速增加[1－2]。由此产生的数据具有丰富内涵，隐藏着很多生物学知识。如何充分利用这些数据，通过合理分析和处理揭示其内涵，获得对人类有意义的信息，为生物学科研工作者带来挑战。

2 生物信息学与医学统计学概述

2.1 生物信息学

包含基因组信息获取、处理、存储、分配、分析和解释的所有方面，是基因组学研究不可分割的一部分；是当下自然科学和技术科学领域中“基因组”、“信息结构”和“复杂性”这 3 个重大科学问题的有机结合[3－5]。生物信息学研究是为了揭示基因组信息结构的复杂性及遗传语言的根本规律，人类在认识自身的基础上可以丰富和发展现有的生物学和信息科学，推动学科群发展，使其成为自然科学中多学科交叉的新领域。

2.2 医学统计学

2.2.1 概述 从 20 世纪 20 年代起，统计学理论与方法日益广泛地被生物医学研究工作者所应用。

〔收稿日期〕 2019－11－12

〔作者简介〕 刘壮，博士研究生，编辑，发表论文 10 余篇。

随着流行病学、基因组学、蛋白质组学、代谢组学等学科迅猛发展，促使统计学与这些学科的交叉融合，对医学统计学研究人员提出很多实践中的新课题。为解决这些新课题，统计学家在对经典统计理论研究和认识的基础上不断探索和发展统计新理论和新方法。医学统计学研究内容主要包括 3 个方面：统计设计、统计分析和其他复杂分析方法。

2.2.2 统计设计 包括对资料收集、整理和分析全过程的设想和安排。在设计前，研究者必须明确的重要问题包括研究目的、研究总体、研究对象、研究内容、样本量、干预措施和研究结果等。在研究设计的构思过程中还应注意几个关键问题，例如抽样方法、控制偏倚和设置对照方法等。

2.2.3 统计分析 主要包括统计描述和统计推断两个部分。统计描述是指用合适的统计图表或统计方法对数据资料的分布状态、数量特征和随机变量之间关系进行估计和测定。统计推断是指在一定的可信程度下由样本信息推断总体特征，包括由样本统计指标（统计量）来推断总体相应指标（参数），即参数估计；由样本差异来推断总体之间是否可能存在差异，即假设检验。

2.3 其他复杂分析方法

在理论统计研究方面，涉及各种概率分布研究、分布偏差的有效性推定以及综合评价方法与理论的研究；在应用统计研究方面，涉及综合评价方法及其应用、统计预测理论与模型研究、各种多元统计方法及其应用的研究、生存时间与生存质量的研究、计算机辅助诊断与治疗模型的研究等。对于这些方面，医学统计学都有相应统计分析方法。

3 统计学方法在生物信息学分析中的应用

3.1 概述

生物信息学中的许多分析方法基本原理都是医学统计学方法的应用和拓展[6-7]。目前生物信息学中常见的问题有序列相似性分析、基因表达分析、基因转录调控网络分析和序列结构与模式识别分析等，本文将介绍这 4 类问题中统计方法的应用情况。

3.2 序列相似性

3.2.1 概述 在分子生物学研究中，对于待研究的碱基序列或由此翻译得到的氨基酸序列，往往需要在数据库搜索到具有一定相似性的同源序列，以推测该未知序列可能属于哪个基因家族，具有哪些生物学功能。序列比较结果一般要经过统计学检验才能判断是否具有显著意义[8]。

3.2.2 Monte Carlo 仿真法 将序列中的符号随机改变后再在同样条件下计算新的配准得分，重复约 100 次后计算样本配准得分的均值和标准差，常被用来判断一对序列配准得分值的统计显著性。在随机序列配准积分符合正态分布的假设下，结果显著性由配准得分高于均值多少个标准差的数目（Z 值）决定。当 Z 值为 3.1、4.3 和 5.2SD 单位时，配准积分的随机出现概率分别是 10^{-3}、10^{-5}和 10^{-7}。通常认为当 Z 值 >5SD 时，两个被比较的序列在进化上相关；当 Z 值在 3 ~5SD 之间时，如果两者在其他方面有相类似的证据可表明两者同源；当 Z 值 <3SD 时，表示两者不同源。

3.2.3 Karlin－Altschul 公式 由于各得分随机变量是在大量分值数据中的最大值（最优配准），正态性假设不尽合理，因此 Karlin 和 Altschul 提出计算 BLAST 得分显著性的 Karlin－Altschul 公式。Vingron 和 Watterman 将此公式推广为适用于计算局部配准得分统计显著性的公式，将序列长度作为其一个参数。对两个序列 a、b，BLAST 发现的高分区匹配域称为 HSPs（high scoring pairs）$a_i \cdots a_{i+k}$ 与 $b_j \cdots b_{j+k}$。最佳 HSP 得分 H（a，b）超过阈值 t 的概率为：

$$P(H(a, b) > t) \approx 1 - e^{-rnmpt} \quad (1)$$

式中 r 和 p 可以通过解一个方程或直接计算得到，m、n 分别是两个序列的长度。式（1）反映 HSPs 得分高于阈值 t 的数目近似为 Poisson 分布。

3.2.4 非重叠局部亚优化配准（Non－overlapping Local Suboptimal Alignment，NOLSA） 那些使局部 Smith－Waterman 配准的期望分值随着被比较序列的长度而呈对数关系增长的罚分称为强 *gap* 罚分。在强 *gap* 罚分的情况下，Karlin－Alts-

chul 公式近似适用于局部配准分析。Waterman 和 Eggert 提出 NOLSA 算法，其中任何一对已经在一种配准中使用过的残疾不再在接下去的较小得分的配准中使用。此算法在每次进行新的次优配准时不必重新计算整个动态规划矩阵，只需重做上一次配准的一个领域，得到的次优配准间的依赖性较低。最优 NOLSA 是 Smith－Waterman 配准。记 ω（t）表示分支不小于阈值 t 的 NOLSAs 数目，可以用 Waterman－Eggert 算法计算 NOLSAs，直到第 1 个 NOLSA 分值 < t 来计算 ω（t）。Karlin－Altschul 公式近似适用于整个区域中的局部配准分析时需要满足 3 个条件：（1）ω（t）近似 Poisson 分布（均值 $\lambda_{n,m}$（t）$=E$（ω（t））；（2）E（ω（t））$=\gamma p^t$，据此可用仿真计算 r、p，此处不需要分布尾部的大量数据；（3）$\gamma = rmn$，据此从一些任意长度的序列得到的 p 与 r 值可以应用于不同长度序列对间配准得分的统计显著性检验，不需要对每对分析的序列重新仿真。在满足这 3 个条件的基础上可以近似计算局部配准的显著性：

$$P(\omega(t) > t) \approx 1 - e^{-rmnpt} \quad (2)$$

3.3 基因表达

3.3.1 概述 随着生命科学进入后基因组时代，基因芯片技术所面临的挑战早已不再是基因表达芯片本身，而是在于发展实验设计方法以对基因表达进行时空全面探索[9]。数据分析与挖掘对其来说是最大挑战。基因芯片表达实验产生海量数据，隐藏着丰富信息，通过数据统计或可视化方法可以发现新的知识。聚类分析是目前运用最多的一种表达数据分析方法。一块基因芯片上往往载有成千上万个基因，一次实验可同时检测这些基因的表达情况。应用同一种芯片在不同条件下（如不同时间、细胞等）进行基因表达实验，搜集表达数据，将原始数据放在一起，生成一个数据表格。表格每一行代表一个基因，每一列代表在不同实验条件下得到的基因表达强度。表格中每一行数据可作为一个向量，聚类分析是将这些向量按照相似程度进行归类。

3.3.2 分层聚类分析[10-11] 在分层聚类情况下，数据被看作是一种二元树结构，在最高层上所有数据同属于一个类。其原理与树的分叉结构相似，类被一分为二，相似的类被保留在同一个子类中，不相似的类则被分开。在进行聚类分析时，从类的每个元素出发将类的集合分为只含有两个类的一组二元类对合集。每个时间中一个类对被合二为一，这样类的数目就减少一个，连续向后进行此过程，最终得到树图的数据分层结构。

3.3.3 K－Means 聚类[12] 在数据划分上不考虑类的分层结构问题。将 R 矩阵的 P 列数组聚为 K 个类，具体方法如下：（1）随机将 R_1，R_2，…，R_P 分配到 K 个类中。（2）计算 K 个类的重心 Y_1，Y_2，…，Y_K。（3）按照由 1 到 P 的顺序计算 R_1，R_2，…，R_P 到重心 Y_1，Y_2，…，Y_K 间的距离，R_i 将分配到距离最近的类中。（4）如果 R_i 被分配到一个新的类中，则重新计算两个受影响的类的重心。（5）重复步骤（3），直到不再有新的类划分出现。

3.4 基因转录调控网络

3.4.1 概述 基因芯片表达数据不仅可用于分析基因表达的时空规律、研究基因功能，还可用于分析基因间的相互关系和基因转录调控网络。单一基因表达结果受其他基因影响，而这个基因同时能影响其他基因表达，这种相互影响、制约的关系构成复杂基因表达调控网络。基因调控网络的研究意义在于通过建立基因转录调控网络统计模型，对某个物种或组织的全部基因的表达关系进行整体分析和研究，分析基因间相互作用。

3.4.2 布尔网络模型 一种以有向图为基础的离散系统，是基因调控分析中最简单的一种模型。在此模型中每个基因只有两种状态，“开”表明基因转录表达，形成基因产物；“关”则表明基因未转录。基因间的相互关系可表示为：

$$\text{A and not B} \longrightarrow \text{C} \quad (3)$$

即如果基因 A 表达，而且基因 B 不表达，则基因 C 表达，其网络图，见图 1。在布尔网络模型中各个基因状态的集合是整个系统的状态，当系统从一个状态转换为另一个状态时，各基因下一时刻的状态由其连接输入机器布尔规则确定。布尔规则用“真值表”的形式表示，当基因 A 和基因 B 处于不

同状态时，基因 C 的状态随之发生变化，见表 1。

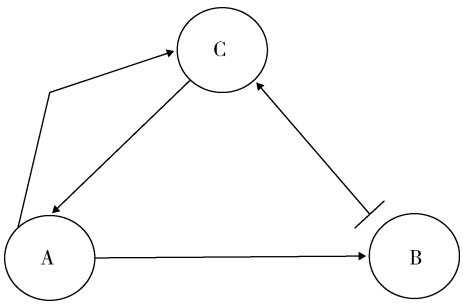

图 1 布尔网络模型

表 1 基因 C 真值

A	B	C
0	0	0
0	1	0
1	0	1
1	1	0

3.4.3 线性组合模型 一种连续网络模型，在此模型中假设基因之间的相互作用是线性的，一个基因的表达值是若干个其他基因表达值的加权和。线性组合模型可表示为：

$$X_i(t+\triangle t) = \sum W_{ij} X_j(t) \quad (4)$$

其中 $X_i(t+\triangle t)$ 是基因 i 在 $t+\triangle t$ 时刻的表达水平，$X_j(t)$ 是基因 j 在 t 时刻的表达水平，为 W_{ij}代表基因 j 的表达水平对基因 i 的影响。在这种基因相互关系表达形式中还可以增加其他数据项，以模拟基因调控的真实情况。

3.5 序列结构与模式识别

结构复杂的蛋白质实际上是由一些相同或不同的结构域缔结而成，每一结构域承担一定功能，各结构域协同作用体现了蛋白质总的生物学功能。测定大量的蛋白质结构可简化为对数量、残基数目较少的结构域结构测定，了解它们如何组装成完整的蛋白质，需要发展新的检索结构域的模式匹配方法。频率表法最先用于核酸序列特殊信号的模式识别，随后逐渐应用于蛋白质结构域的模式匹配分析中。由于蛋白质的结构域通常由几十个或几百个残基组成，属于同一类结构域的序列的类似性可能很小。结构域保守区决定了结构域的同源，因此其存在确定了结构域的存在，可以用结构域的保守顺序直接分析蛋白质与蛋白质超家族的类似性，增加检测敏感性。

4 结语

作为连接生命科学和信息科学的新兴学科，生物信息学发展前景广阔。而统计学作为生物信息学分析的重要工具，可以探查和提取数据之间的因果关系，揭示数据内涵，从而获得更多有价值的信息。本文通过介绍序列相似性分析、基因表达分析、基因转录调控网络分析和序列结构与模式识别分析中统计学方法的应用，为科研人员学习系统的生物分析技术提供理论依据。

参考文献

1 International Hap Map Consortium. A Haplotype Map of the Human Genome [J]. Nature, 2005, 437 (7063): 1299 - 1320.
2 孙啸. 生物信息学基础 [M]. 北京: 清华大学出版社, 2005.
3 Botstein, David, Neil Risch. Discovering Genotypes underlying Human Phenotypes: past successes for mendelian disease, future approaches for complex disease [J]. Nature Genetics, 2003 (33): 228 - 237.
4 Mortazavi A, Williams BA, McCue K, et al. Mapping and Quantifying Mammalian Transcriptomes by RNA - Seq [J]. Nature Methods, 2008, 5 (7): 621 - 628.
5 Anders S, Huber W. Different Expression Analysis for Sequence Count Data [J]. Genome Biology, 2010, 11 (10): R106.
6 马世琪. 生物信息中的统计模型及其仿真与应用 [D]. 成都: 电子科技大学, 2016.
7 郭乐乐. 统计聚类在生物信息分析中的应用 [D]. 兰州: 兰州大学, 2014.
8 Yang Jian. Genome Partitioning of Genetic Variation for Complex Traits Using Common SNPs [J]. Nature Genetics, 2011, 43 (6): 519 - 525.
9 黄德双. 基因表达谱数据挖掘方法研究 [M]. 北京: 科学出版社, 2009.
10 杨小兵. 聚类分析中若干关键技术的研究 [D]. 杭州: 浙江大学, 2005.
11 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述 [J]. 计算机应用研究, 2007, 24 (1): 10 - 13.
12 Yi Hong, Sam K. Learning Assignment Order of Instances for the Constrained K - means Clustering Algorithm [J]. IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, 2009, 39 (2): 568 - 574.