计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):46−54 [doi: 10.15888/j.cnki.csa.008279]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

肿瘤突变印迹的求解算法比较①

何小雨[1,2], 韩鑫胤[1,2], 牛北方[1,2]

[1](中国科学院 计算机网络信息中心, 北京 100083)
[2](中国科学院大学, 北京 100049)
通讯作者: 牛北方, E-mail: bniu@sccas.cn

摘　要: 细胞受到致癌因子刺激时引起突变, 突变过程使基因组发生具有一定模式的改变, 称为突变印迹. 突变印迹分析是阐明致癌因子的致癌机制及驱动癌症发展的一项重要任务, 将为肿瘤早期诊断和个体化治疗提供新的依据和选择. 下一代测序技术的突破和发展使得海量体细胞突变被识别, 从而使从大规模基因组中挖掘突变印迹成为可能. 本文详细解释了突变印迹识别问题的数学模型, 介绍了求解方法和重要参数, 系统全面地比较了主流算法和软件, 指明了突变印迹提取的注意事项. 最后, 对该领域的未来发展趋势进行了探讨.

关键词: 体细胞突变; 突变印迹; 非负矩阵分解; 期望最大化; 线性回归

引用格式: 何小雨,韩鑫胤,牛北方.肿瘤突变印迹的求解算法比较.计算机系统应用,2021,30(12):46–54. http://www.c-s-a.org.cn/1003-3254/8279.html

Comparison of Algorithms for Tumor Mutational Signature

HE Xiao-Yu[1,2], HAN Xin-Yin[1,2], NIU Bei-Fang[1,2]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China)
[2](University of Chinese Academy of Sciences, Beijing, 100049, China)

Abstract: Mutations are induced when cells are stimulated by carcinogens. The mutational process causes a certain pattern of changes in the genome, which are called mutational signatures. Mutational signature analysis is an important task in clarifying the carcinogenic mechanism of carcinogens and driving the development of cancer research, and it will provide new insights and options for early tumor diagnosis and individualized treatment. The breakthroughs and developments of next-generation sequencing have led to the identification of massive somatic mutations, making it possible to mine mutational signatures from large-scale genomes. This study elaborates the mathematical model for mutational signature identification and introduces alternative methods and important parameters. It systematically and comprehensively compares mainstream algorithms and software and specifies the precautions for mutational signature extraction. Finally, it forecasts the future development trend of this field.

Key words: somatic mutation; mutational signature; nonnegative matrix factorization; expectation maximization; linear regression

癌细胞的基因组携带体细胞突变 (somatic mutation). 体细胞突变指发生在除生殖细胞之外的人体细胞中的DNA 结构上碱基对的改变, 其过程贯穿个体的整个生命周期. 体细胞突变的诱因主要有物理因素 (如射线)、化学因素 (如抗生素、烟草) 以及生物因素 (如细菌或病毒基因的融合)[1]. 如图 1 所示, 从受精卵第一次

① 基金项目: 中国科学院战略性先导科技专项 (B 类)(XDB38040100)
Foundation item: Strategic Priority Research Program of the Chinese Academy of Sciences (Category B) (XDB38040100)
收稿时间: 2021-03-08; 修改时间: 2021-03-31, 2021-04-13; 采用时间: 2021-05-07

分裂开始, 细胞受到某种致癌因子的刺激而发生突变. 突变过程可能会引起 DNA 损伤和修复, 并产生单核苷酸替换 (single nucleotide substitutions)、短插入和删除 (short insertions and deletions)、结构变异 (structure variants) 和染色体拷贝数变化 (copy number variation) 等[2]. 各致癌因子之间相互独立, 在体细胞基因组上留下独特的记号, 随着个体生命的延续, 体细胞突变不断叠加累积, 最终使细胞脱离控制, 形成生长增殖不受限制的癌细胞. 生物医学研究中普遍认为, 细胞在正常生长分裂过程中产生的突变在基因组上是随机分布的, 而由特定致癌因子导致的突变则具有一定的模式. 这种模式能够反映细胞癌变过程中曾暴露于哪些致癌因子以及在其中的暴露程度. 研究中将基因组中由致癌因子引起的这种特有突变模式叫做突变印迹 (mutational signature)[3].

早在 DNA 双螺旋结构被发现之前, 已有临床研究发现连续的紫外线辐射会加速细胞增殖的相对速度, 从而明确了紫外线的过量辐射是皮肤癌的一大诱因[4]. 测序技术的出现尤其是下一代测序技术 (Next Generation Sequencing, NGS) 的蓬勃发展, 迅速推动了分子生物学和生物信息学的研究进展[5]. 研究显示, 人的基因组序列中存在大约 400 万个突变位点, 全外显子组序列中可以检测到大约 6 万–8 万个突变位点[6]. 大多数癌症中的体细胞突变属于“乘客”突变 (passenger mutations), 即不导致细胞癌变, 在癌症发展的过程中也不会被正向选择. 仅有少数的突变是“驱动”突变 (driver mutations), 即赋予细胞生长优势并且帮助癌细胞增殖[7].

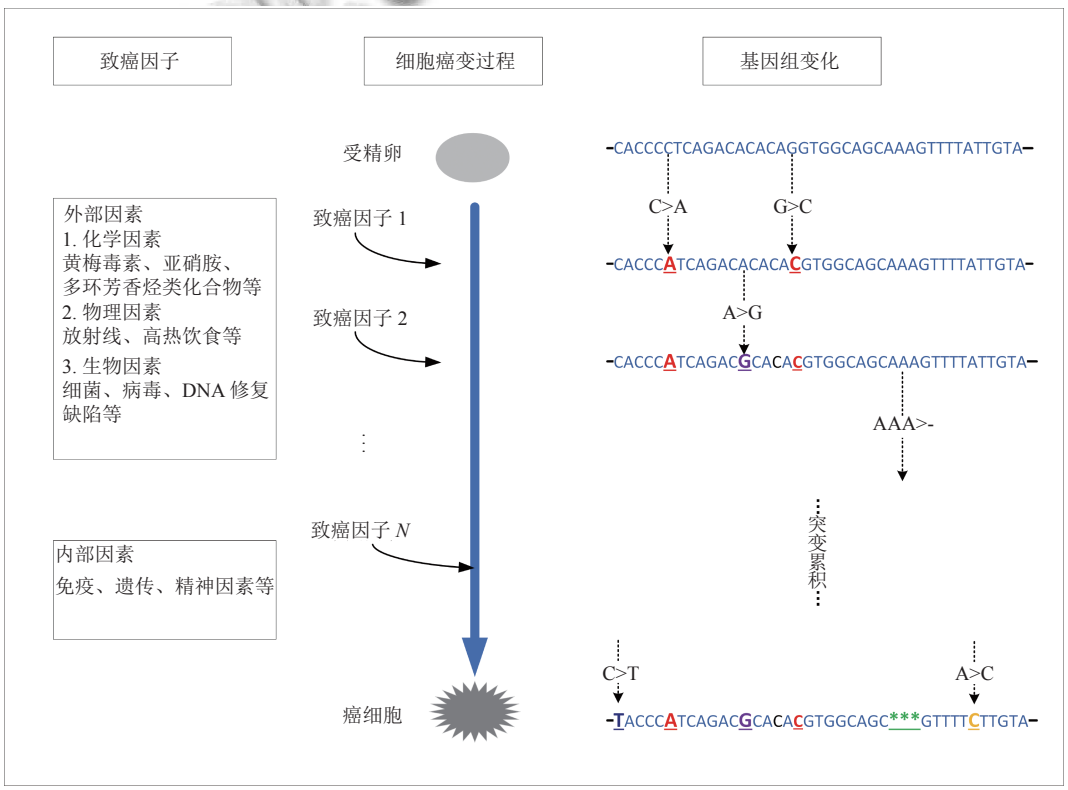

图 1 突变过程在癌症基因组中留下的特征性印记

长久以来, 对体细胞突变的研究方向局限在驱动基因 (driver genes) 上, 如 *TP53*[8]. 然而, “乘客”事件可能不是癌症发展的原因, 却一定是细胞癌变过程中的产物. 因此“乘客”突变携带并记录了 DNA 损伤和修复的丰富历史证据, 这些证据在癌症的发生发展中起关键作用. 因此, 探寻突变过程时应充分考虑“乘客”突变. 突变印迹分析将发生在基因组上的所有突变位点纳入考虑, 是对癌症基因组研究的重要补充.

突变印迹的概念由 Alexandrov 等人提出的[3], 他们在 21 个乳腺癌的全基因组单碱基替换突变类型中计算并提取了 5 个特殊的突变模式, 这些单基因替换突变并未进行“驱动”和“乘客”的区别, 而是将他们全部

作为分析对象. 研究中还尝试解释了其生物学机制, 如以 TpCpX 三核苷酸处的 C>T, C>G 和 C>A 替换为主要特征的突变印迹是由 5-甲基胞嘧啶作用自发的内源性突变过程导致的, 是癌症样本中绝大多数的突变类型, 并且以这种显性的突变形式存在于约 10% 的雌激素受体 (Estrogen Receptor, ER) 阳性乳腺癌患者中[2]. 虽然当时并未对计算出的 5 个印迹进行完全的生物学解释, 但该研究已经对潜在的突变机制形成了一些见解并提供了研究的思路. 2013 年, Alexandrov 等人又开展了一项涉及 30 个癌种, 7042 个样本的研究, 在 4938362 个单碱基替换中解析出 21 个经过生物学验证的突变印迹[3]. 2018 年, 该研究团队将研究数据集继续扩大到 23829 个癌症样本的 4729 690 个体细胞突变集合上, 同时纳入了更多的突变类型: 单碱基替换, 双碱基替换, 短插入删除和结构变异. 更大规模的数据集中发现了新的突变印迹: 49 种单碱基替换印迹, 11 种双碱基替换印迹, 17 个短插入删除印迹以及 11 个结构变异的聚类[9]. 这些印迹既验证了紫外线、烟草、酒精等与癌症发生相关联的外部因素, 一些新的印迹也被证实与独特的临床特征相关, 这说明突变印迹的分析很可能成为靶向治疗中新的潜在生物标志物[10].

虽然一些突变印迹反映的致癌因子 (如紫外线照射、烟草) 可以通过跟踪调查和统计来识别, 但是一种正式的数学方法可以提取人类难以察觉的更微妙的元素和较弱的信号, 同时还需要评估该元素或信号在癌变过程中的比例. 并且随着基因组计划的完成, 大规模的基因组测序数据也对突变印迹问题的求解提出了考验. 目前, 已有多个基于 NGS 的识别突变印迹的算法和软件, 但对该问题的求解在参数设置等方面尚未达成共识. 据调研, 目前缺乏关于体细胞突变印迹分析算法和软件的比较, 并且随着更多癌症突变位点的检出, 突变印迹分析将获得很大的挖掘空间. 因此, 开展该领域的相关研究, 清晰详细地介绍和讨论是十分必要的.

总体来讲, 本文的主要研究内容包括:

(1) 突变印迹问题及其数学模型阐述.

(2) 突变印迹提取算法及评价.

(3) 突变印迹提取软件的其他功能介绍.

(4) 比较总结实验结果并提出新的解决方案.

1 突变印迹的数据模型

突变印迹问题可以描述为: 从复杂的体细胞突变信号中寻找独立致癌因子使基因组发生的特有的改变模式. 癌细胞基因组的改变可能是多个致癌因子引发突变的累积, 原始致癌因子使基因组发生改变的程度也不相同. 因此, 可将突变印迹问题抽象为盲源分离问题 (Blind Source Separation, BSB)[11].

盲源分离问题是研究在系统的传递函数、源信号的混合系数及概率分布未知的情况下, 利用源信号之间相互独立这一微弱的已知条件, 如何从一组复杂的混合信号中分离出独立的不可观测的源信号. 盲源分离作为阵列信号处理的一种新技术, 允许有意义地学习对象的不同部分, 近几年来受到广泛关注.

盲源分离问题在突变印迹分析中的应用可以表述为算法 1.

算法 1. 突变印迹分离算法

1. 计算体细胞突变信号中的最优突变印迹的组合, 以表示在癌症的发展过程中每个独立突变过程的累积;
2. 计算每个印迹在每个独立癌症基因组的体细胞突变中的比例, 表示印迹对应的致癌因子对癌变过程的贡献度.

算法中原始体细胞突变信号常采用结合上下文的三碱基结构表示. 基因组是由腺嘌呤 (A)、胸腺嘧啶 (T)、鸟嘌呤 (G)、胞嘧啶 (C) 组成的序列. 由于基因组正负链上的碱基以互补配对原则, 即 A 与 T 配对, G 与 C 配对形成碱基对, 则基因组上可能出现的单核苷酸替换情况有C_4^2种 (图 2), 其中:

(1) C>A: 代表 C>A 和 G>T 两种单核苷酸替换方式;

(2) C>G: 代表 C>G 和 G>C 两种单核苷酸替换方式;

(3) C>T: 代表 C>T 和 G>A 两种单核苷酸替换方式;

(4) T>A: 代表 T>A 和 A>T 两种单核苷酸替换方式;

(5) T>C: 代表 T>C 和 A>G 两种单核苷酸替换方式;

(6) T>G: 代表 T>G 和 A>C 两种单核苷酸替换方式.

将这 6 种替换方式表示为如式 (1) 所示的单碱基替换字典 V:

$$V=\{[\mathrm{C>A}],[\mathrm{C>G}],[\mathrm{C>T}],[\mathrm{T>A}],[\mathrm{T>C}],[\mathrm{T>G}]\} \quad (1)$$

然后, 将其上游 (5′端)、下游 (3′端) 各一个碱基作为其上下文 (如图 2 中标识), 三碱基结构表示为式 (2):

$$T=\{5'-XvY-3' \mid X,Y\in\{\mathrm{A,C,G,T}\},\mathrm{v}\in\mathrm{V}\} \quad (2)$$

显然, 单碱基上下文结构的表示方法有$4\times C_4^2\times 4=96$ 种. 则单个基因组上的体细胞突变谱可表示为如式 (3) 所示的向量:

$$T_n=[T_{1n},T_{2n},T_{3n},\cdots,T_{kn}]^{\mathrm{T}},\ 1\le k\le 96 \quad (3)$$

其中, T_{kn} 表示与 T_n 关联的突变过程 (第 n 个致癌因

子) 引起的第 k 个突变符号的频率. 因此:

$$\sum_{k=1}^{96} T_{kn} = 1;\ 0 \le T_{kn} \le 1, 1 \le k \le 96 \tag{4}$$

将一个癌症队列中每个患者的体细胞突变表示为如式 (3) 所示向量, 则该队列的突变目录即可表示为:

$$M_g = \left[m_{1g}, m_{2g}, \cdots, m_{kg}\right] \tag{5}$$

其中, 突变目录的每个元素可以近似地认为是使正常细胞发展为肿瘤细胞的潜在突变过程的特征的线性叠加, 且每个特征通过在相应过程中的暴露程度来加权, 如式 (6):

$$m_{kg} = \sum_{n=1}^{N} s_{kn} \times e_{ng} \tag{6}$$

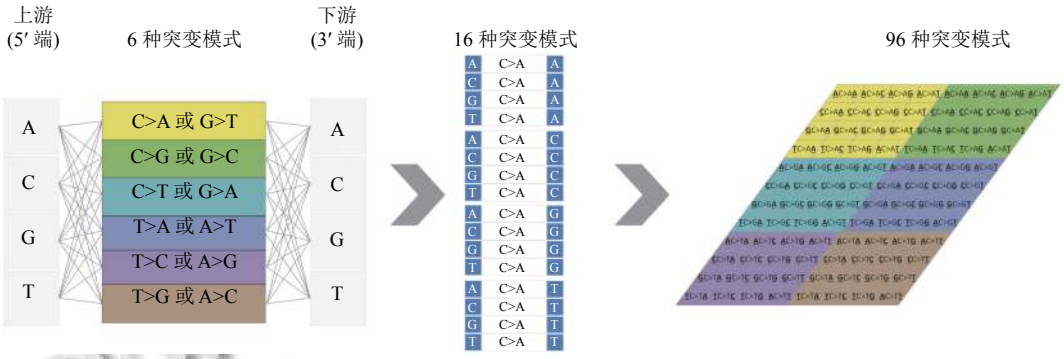

图 2 从体细胞突变列表中构建 96 种突变模式

因此突变印迹可通过将 M 表示为两个较小的矩阵实现. 即:

$$M \approx SE \tag{7}$$

其中, M 是研究队列中 g 个基因组的 96 个突变类型的突变频率:

$$M_{k\times g} = \begin{bmatrix} m_{11}, m_{12}, \cdots, m_{1g} \\ m_{21}, m_{22}, \cdots, m_{2g} \\ \vdots \\ m_{k1}, m_{k2}, \cdots, m_{kg} \end{bmatrix} \tag{8}$$

而 $S_{k\times n}$ 是该队列基因组由 n 个致癌因子导致的特定的突变模式:

$$S_{k\times n} = \begin{bmatrix} S_{11}, S_{12}, \cdots, S_{1n} \\ S_{21}, S_{22}, \cdots, S_{2n} \\ \vdots \\ S_{k1}, S_{k2}, \cdots, S_{kn} \end{bmatrix} \tag{9}$$

$E_{n\times g}$ 表示 g 个基因组在 n 个致癌因子中的暴露程度:

$$E_{n\times g} = \begin{bmatrix} e_{11}, e_{12}, \cdots, e_{1g} \\ e_{21}, e_{22}, \cdots, e_{2g} \\ \vdots \\ e_{n1}, e_{n2}, \cdots, e_{ng} \end{bmatrix} \tag{10}$$

此外, 突变印迹因其代表的自然意义而具有两个特征: (1) 研究对象的非负性. 一个癌症队列中所有基因组体细胞突变目录 M 表示的是每个样本的每种突变类型的突变频率, 分解矩阵 S 表示突变过程的特征, 系数矩阵 E 表示突变过程在基因组上的强度; (2) 研究目标是期望从体细胞突变目录中提取具有生物学意义的特征, 即致癌因子在基因组上留下的特殊记号.

2 算法分类和比较

根据对式 (7) 的求解方法, 可以将突变印迹问题分为 3 类: 一是非负矩阵分解 (Nonnegative Matrix Factorization, NMF) 的方法; 二是期望最大化 (Expectation Maximization, EM) 方法; 三是线性回归 (Linear Regression, LR) 方法. 按照是否能够发现新印迹可以将突变印迹问题分为允许新印迹发现的“提取”方法和对已知印迹的“拟合”方法.

2.1 求解方法的比较

2.1.1 NMF 方法

NMF 算法是由 Lee 和 Seung 于 1999 年提出的一种矩阵分解方法, 它使分解后的所有分量均为非负值,并且同时实现非线性的维数约减[12]. NMF 作为从各种类型的高维生物数据中提取有意义成分的一种强大技术屡次脱颖而出. 此外在其他领域也有 NMF 方法的成功应用, 如“鸡尾酒会”问题, 人脸识别问题等[13]. 由

于上述数学模型中矩阵的内在非负性, 使得 NMF 特别适合于突变特征推断问题. NMF 也是第一个用来尝试分析突变印迹问题的算法. 如表 1 所示, 目前已有如 SigProfiler、SomaticSignatures、sigfit 和 MutationalPatterns 等多个软件基于 NMF 算法来解决印迹分解问题.

SigProfiler 方法通过找到矩阵 S 和 E 来准确地提取 N 个突变印迹, 同时解决由式 (7) 导出的非凸优化问题. 该方法选择矩阵范数作为 Frobenius 重构误差:

$$\min_{S\ge 0,E\ge 0}\|M-SE\|_F^2 \tag{11}$$

表 1 突变印迹分析软件汇总

软件名	分析方法	计算原理	运行环境	输入数据	印迹个数	可视化
SigProfiler	提取	NMF	Matlab	突变目录矩阵	×	√
EMu	提取	EM	Python	突变目录矩阵	√	×
SomaticSignatures	拟合	PCA; NMF	R	突变识别格式	×	√
pmsignature	拟合	EM & NFM	R Web Page	突变列表	×	√
signeR	拟合	NMF	R; C++	突变识别格式	√	√
DeconstructSigs	拟合	LR	R	突变列表	×	√
mutagene	提取	NMF	WebPage	突变注释格式	×	√
Mutalisk	拟合	LR	WebPage	突变识别格式	×	√
sigfit	提取拟合	NMF	R	突变识别格式	√	√
Helmsman	提取	NFM; PCA	Python	突变识别和注释格式	√	√
MutationalPatterns	提取	NMF	R	突变识别格式	√	√

以 SigProfiler 为代表的 NMF 算法具体步骤如算法 2.

算法 2. 以 SigProfiler 为代表的 NMF 算法

1. 初始化随机非负矩阵 S, E;
2. 将初始突变目录矩阵降维, 将所有突变类型中占比≤1% 的突变类型删除, 得到矩阵 M';
3. 迭代:
(1) 对矩阵 M'进行蒙特卡洛自举重采样 (Monte Carlo bootstrap resampling), 得到矩阵 M'';
(2) 乘法更新算法应用于 M'', 得到使式 (11) 中的 Frobenius 范数最小的 S 和 E;
4. 对 S 划分聚类, 得到 N 个簇;
5. 将每个簇中的 s 归一化, 得到 S 的 N 个向量;
6. 求暴露矩阵 E.

基于 NMF 方法提取突变印迹的软件需要指定分解个数 N 作为程序输入, SigProfiler 根据大量实验给出的建议值是:

$$N\in(1\sim\min\{K,G\}-1) \tag{12}$$

在实际操作中, SigProfiler 根据每个 N 值计算模型的总体再现性和 Frobenius 范数误差. 最终需要人工干预选择 N 值, 使得分解的印迹矩阵 S 具有高度的再现性同时显示出低的总体重建误差.

2.1.2 EM 方法

Fischer 等人基于 NMF 方法从基因组内在特性出发, 应用概率模型解决突变印迹问题[14]. 基因组的内在的特性 (如 CpG 双核苷酸的不均一分布, 拷贝数变化) 会影响三碱基序列结构发生突变的可能性, 继而使模型在推断的突变模式上产生偏倚. 理论上, 使用概率模型可在求解过程中充分考虑突变发生的可能性, 能够更准确地分离出真实突变过程的印迹.

将三碱基序列结构的突变可能性表示为非 0 的 k 元组, 其中 O_{kg} 表示基因组 g 上第 k 个突变类型发生突变的可能性:

$$O_g=[O_{1g},O_{2g},O_{3g},\cdots,O_{kg}];\ 1\le k\le 96 \tag{13}$$

EM 算法的目标是使包含隐变量的数据集的后验概率或似然函数最大化, 进而得到最优的参数估计. 文献 [14] 将突变印迹分解问题重构为一个如式 (14) 所示的概率模型. 其中突变目录矩阵 (M) 分布为独立的泊松随机变量, 其元素由印迹矩阵 (S) 与暴露矩阵 (E) 乘积确定, 通过期望最大化算法对 S 和 E 进行估计.

$$P\left(M_g\mid E_g,O_g,S\right)\equiv\prod_{k=1}^{K}Pois\left(m_k\mid O_k\sum_{j=1}^{x}s_kv_{ng}\right) \tag{14}$$

算法的具体执行过程如算法 3.

算法 3. 以 EMu 为代表的 EM 算法

1. 猜想模型参数 $S^{(0)}$;
2. 迭代
(1) 给定当前猜测参数 $S^{(k)}$, 得到曝光估计值$\hat{E}$;
(2) 使用$\hat{E}$更新下一次迭代的参数估计值 $S^{(k+1)}$;
(3) 当 $P(M|S)$ 收敛到极大值时, 迭代结束;
3. 比较不同 N 下的数据可能性, 确定突变过程数量.

值得注意的是, 虽然 EMu 是建立在对 NMF 的有效替代解释的基础上, 该解释将 NMF 视为对特定问题的 EM 应用, 但 EMu 的新概念和优点并不是固有的 EM 范例特性, 也可以通过其他方法进行同化的显式增强. 另一方面, EMu 对初始条件的敏感度与常规 NMF 相同. 尽管如此, EMu 成功地利用了突变印迹推断的概率形式来解决以前未曾探索过的方向, 即结合了基因组的内在特性和肿瘤特定的突变可能性并确定了突变印迹的个数.

2.1.3 LR 方法

线性回归 (Linear Regression, LR) 指通过对大量的观测数据进行处理, 从而得到比较符合事物内部规律的数学表达式[15]. 在 NMF 方法发现了一些可解释的突变印迹的基础上, 印迹提取问题可扩展为新的描述: 对某个癌症队列的研究不再需要发现新的印迹, 而重点在于得到肿瘤中存在的、可解释的印迹以及其对癌症发展的作用程度. 因此, 突变印迹提取转变为对已知突变印迹分布的拟合, 即对线性方程 (15) 的求解.

$$M = SE + R \tag{15}$$

由于线性回归方法对先验知识的依赖, 尤其依赖对已知突变印迹的个数选择和组合, 导致其在实际应用中十分受限, 如目前已知的突变印迹集尚不能完全地解释癌变过程, 个别突变印迹没有得到完备的生物学解释等. 同时研究中也指出该方法的准确性偏低, Maura 等人在 2019 年的研究中发现, 使用同样的突变目录 (M) 情况下, 非负矩阵分解算法提取的突变印迹大部分能够被验证其所代表的生物学意义, 而使用线性回归方法拟合的突变印迹大多为“无特征”或“平坦”印迹, 即 6 种突变模式的频率分布相对均匀, 无明显差异[16].

2.2 “提取”算法和“拟合”算法

表 1 中的软件和算法分别从不同角度出发来解决印迹问题, 根据算法特征及必需的输入数据, 可将其分为印迹“提取”算法和“拟合”算法. 基于 NMF 的“提取”算法以突变目录矩阵 M 和分解个数 N 为输入, 求解印迹和暴露矩阵, 因此分解出的印迹矩阵 S 中可能会出现新的印迹. 与提取方法不同, 拟合方法以目录矩阵 M 和已知的印迹矩阵 S 为输入, 将 M 中潜在的印迹拟合为 S 的线性表达.

“提取”的方法的优势是不依赖先验知识 (已知的突变印迹), 同时允许提取出新的突变印迹. 该算法也存在局限性. 首先, 同时发生的多个独立的突变印迹可能会被合并为一个印迹. 其次, 对于非常复杂的印迹可能会因其较小的贡献度而拆分为两个或多个印迹.

反之, “拟合”的方法依赖大量先验知识, 如分析的癌种有哪些致癌因子, 这些致癌因子的突变印迹分别是什么等. 现有公开发布的突变印迹总计有 81 种, 且 50% 尚未得到生物学解释, 突变目录矩阵本身的大小以及疾病类型乃至分型都会对印迹个数和组合的选择产生影响. 由于其对已知突变印迹的依赖, 因而不能发现新的印迹. 此外, 当主观性较强的先验知识被输入时, 很可能导致过拟合现象, 即夸大某个印迹在该癌种发生发展中的权重. 反之, 抛弃先验知识的限制, 将全部已知的突变印迹作为输入, 则会导致特异性突变印迹的渗透, 即少量样本的突变印迹分配至整个队列的样本中或拟合到并未在癌种的发生发展中起作用的印迹上.

同时, 两类算法也存在共性问题, 当不同的突变印迹的组合可解释同一个突变目录矩阵时, 印迹提取就会变得不明确; 当队列中存在少量异质性较高样本时, 其突变印迹因贡献度不高而被过滤掉, 从而掩盖队列的异质性.

3 突变印迹分析软件的其他功能

在实际应用中, 除了对式 (7) 的求解外, 突变目录矩阵的构建、突变印迹及暴露矩阵的可视化、运行环境等问题在生物信息学领域也是广受关注的需求.

获取突变上下文是构建突变目录矩阵的关键. 基于 NGS 的突变识别软件如 VarScan2[17]、Strelka2[18]、SomaticSniper[19] 等给出的突变并不包含上下文信息. 有些软件如 VarDict[20] 在 INFO 列使用“LSEQ”和“RSEQ”分别给出突变上下文, 但大多数的软件不能提供此信息, 需要重新计算. 可解决的方案是利用突变位置从参考基因组中获取上下文. 但是人类的参考基因组有超过 30 亿个碱基, 分别计算一个基因组中上百万个突变位点的上下文需要一定的时间消耗. 此外, 通过突变注释软件 (如 Oncotator[21]) 能够获得上下文碱基序列, 然而这就需要在计算突变印迹前先对突变列表进行注释, 增加了构建突变印迹分析流程的复杂性. 因此很多软件在内部增加了计算上下文的过程, Pedersen 等人开发了对 VCF 快速处理的 Python 程序 Cyvcf2, 可实现快速的 VCF 文件处理[22].

突变印迹和暴露矩阵可视化也是基因组数据挖掘的重要内容. 除了 EMu 外, 目前用于计算突变印迹的软件都提供了模块化的可视化方法. 这些可视化方法都直接使用各自的分析结果作为输入, 在实际操作中无需分析人员具备专业的绘图知识.

最后, 基因组数据挖掘软件大多依赖 Linux 环境, 需要操作人员具备在 Linux 环境下编译、安装和运行软件的能力. 因此, 除了发布软件包, 多个软件如 PMSignature、MutaGene 等提供了门户网站的计算方式, 这在很大程度上方便了缺少计算机知识的生物医学研究人员, 同时这也是整个生物信息学领域软件的发展趋势.

4 实验分析

4.1 数据说明

本文选择了 21 个宫颈癌的全基因组测序数据, 分别使用基于非负矩阵分解算法的软件 SigProfiler[3] 和基于线性回归方法的软件 Mutalisk[23] 分析突变印迹. 由于基于期望最大化的方法需要关键的先验知识 (基因组序列的复杂性), 本文实验中不包括对 EM 算法的实验.

4.2 实验过程

SigProfiler 的实验过程如下:

(1) 构建体细胞突变目录矩阵;

(2) 将最大提取印迹数设置为 10, 即 N 的最大取值为 10;

(3) 运行 SigProfiler;

(4) 选择结果稳定性高且残差相对较小的突变印迹的个数的最大值 N';

(5) 提取 N'个突变印迹, 构成集合 S;

(6) 计算 S 中每个印迹与公开发表印迹的相似性.

Mutalisk 的实验过程如下:

(1) 合并 21 个基因组的突变识别格式文件;

(2) 选择 COSMIC 的 30 个突变印迹作为先验输入 S;

(3) 运行 Mutalisk;

(4) 选择文献中记载的与宫颈癌相关的印迹作为先验输入 S';

(5) 再次运行 Mutalisk.

4.3 实验结果

图 3 表示了对 21 个癌症患者的体细胞突变目录矩阵进行 NMF 分解时, 不同的突变印个数对应的突变印迹稳定性 (左) 和原始突变目录矩阵的重建误差 (右).

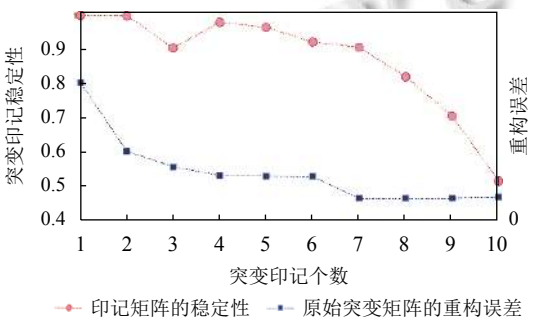

图 3 突变印迹数对应的突变印迹稳定性和重建误差

实验中, 将分解印迹数 N 设置为 [1,10], 对每个 N 的取值分别计算两个指标, 最终选择再现性高且重建误差小的印迹个数. 通过图 3 可以看出, 当突变印迹数 N<7 时, 随着 N 的增大, 再现性逐渐降低 (N=3) 除外, 重构误差逐渐降低, 而在 N>7 时, 再现性显著降低, 重构误差稳定. 因此, 为了同时保证印迹矩阵的稳定性高和原始矩阵的重构误差低, 选择印迹个数为 7. 图 4 表示了具体 7 个突变印迹.

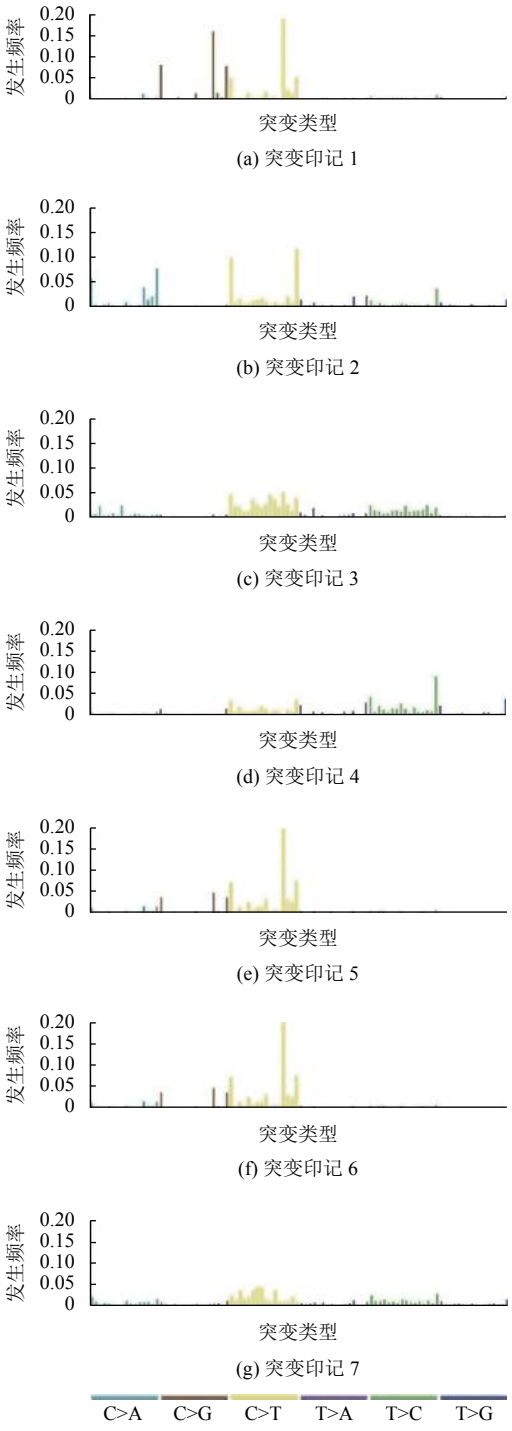

图 4 21 个癌症基因组中提取的 7 个突变印记

为了说明提取的 7 个印迹与该癌种发病之间的关系, 计算了 7 个突变印迹与国际公开数据库 COSMIC (Catalogue Of Somatic Mutations In Cancer) 中 2017 年收录的已知 30 种突变印迹的余弦相似性以及可能的生物学解释 (表 2).

表 3 是 Mutalisk 采用 30 个 COSMIC[24] 的印迹为输入时的结果, 其将 21 个宫颈癌基因组的突变目录拟合到 S5, S2, S13, S8, S1, S28 六个已知印迹中.

当选择已知相关印迹 (S1, S2, S5, S6, S8, S10, S13, S26) 为先验知识时, Mutalisk 给出 5 种拟合结果. 其中选择贝叶斯信息准则 (Bayesian Information Criterion, BIC) 最小的拟合结果, 即 S5, S2, S13, S1, S6 (表 4).

表 2 SigProfiler 提取突变印迹与已知印迹相似性

印迹ID	参考印迹ID	余弦相似性	解释
1	S2	0.758	APOBEC
	S13	0.731	APOBEC
2	S8	0.584	N/A
3	S5	0.781	N/A
	S29	0.776	烟草
4	S5	0.748	N/A
5	S2	0.914	APOBEC
6	S22	0.897	N/A
	S25	0.850	N/A
7	S5	0.844	N/A

表 3 Mutalisk 与已知 30 种拟合印迹情况及相似性

印迹ID	参考印迹ID	余弦相似性	解释
1	S5	0.421	N/A
2	S2	0.209	AID/APOBEC
3	S13	0.112	AID/APOBEC
4	S8	0.112	N/A
5	S1	0.112	5-甲基胞嘧啶脱氨作用
6	S28	0.034	N/A

表 4 Mutalisk 与已知 6 种印迹拟合情况及相似性

印迹ID	参考印迹ID	余弦相似性	解释
1	S5	0.556	N/A
2	S2	0.217	APOBEC
3	S13	0.110	APOBEC
4	S1	0.087	5-甲基胞嘧啶脱氨作用
5	S6	0.030	DNA错配修复缺陷和微卫星不稳定性

4.4 实验总结

通过对比表 2 至表 4 中的余弦相似性可知, 基于非负矩阵分解的软件 SigProfiler 发现的印迹与国际研究公开的印迹相似性均高于基于线性回归的拟合方法. 虽然 SigProfiler 提取的 2, 6, 7 并未得到相应的生物学解释, 但其余 4 个印迹分别与 APOBEC 基因家族和吸烟有关. 既往研究显示, APOBEC 基因家族的突变印迹存在于乳腺癌、宫颈癌、肺癌等多个癌种, 其在慢性炎症条件下的异常表达可能误伤人类本身的基因组[25].

此外, 基于线性回归的拟合方法虽然也拟合到了 APOBEC 基因家族, 但其相似性极低. 当使用先验知识干预时, 相似性有微弱的提升, 但并不具有统计效力. 同时, 先验知识的加入使得拟合结果中过滤掉了印迹 8 和 28, 出现了印迹 6, 而印迹 6 是一个“平坦”印迹, 所以先验知识的加入也没有使得拟合出现可接受的结果. 最后, 该实验验证了在印迹分析问题中, NMF 方法比 LR 方法适用性更强.

5 结束语

本文全面详尽地探讨了体细胞突变印迹分析的相关概念和模型, 并对算法进行了说明, 分类和比较; 阐述了基因组测序数据进行突变印迹提取的注意事项和缺陷. 此外, 使用真实的基因组数据全面而详细地示范了如何在癌症基因组的研究中应用突变印迹分析.

本文介绍的分析框架使用了一些重要的限制和假设来描述基因组上的突变. 因此, 目前已经提取到的突变印迹仍然是数学近似值, 其轮廓可能受到所用数学方法的影响. 从概念和使用的简单性出发, 研究中假设一个印迹与某个致癌因子引起的特定突变过程相关联, 并以均一化的形式来表示. 然而, 不同的数学方法可以发现具有相似性和差异性的印迹, 并且这些印迹已通过多种方式得到证实. 随着突变数量的增加和不同类型突变之间的数量级差异, 单一数学方法可能无法实现准确的印迹分离, 因此结合充分的先验知识, 进一步研究破译和鉴定突变印迹的方法是避免产生在生物学上不可信或难以解释结果的有效手段.

突变印迹的研究已经发现了一些诱导癌症发生发展的原因, 人类癌症中自然发生的突变特征很可能有相当一部分已经被描述出来. 然而, 一些罕见的或者由治疗导致的突变印迹可能还没有被捕捉到, 需要进行彻底的探索. 目前许多最新发现的印迹背后的机制尚未明确, 还有待进一步地实验和理解. 未来, 突变印迹的提取还应该包含更多的突变类型, 无论是致病型还是继发型, 同时也包括由遗传导致的癌症易感基因中的种系突变. 这些印迹背后的诱因对于癌症预防和公共卫生具有重要意义.

参考文献

1 冯娟. 癌症病因新进展: 更多体细胞突变基因被鉴定. 生理科学进展, 2007, 38(2): 173–173. [doi: CNKI:SUN:SLKZ.0.2007-02-024]

2 Nik-Zainal S, Alexandrov LB, Wedge DC, *et al*. Mutational processes molding the genomes of 21 breast cancers. Cell, 2012, 149(5): 979–993. [doi: 10.1016/j.cell.2012.04.024]

3 Alexandrov LB, Nik-Zainal S, Wedge D, *et al*. Deciphering signatures of mutational processes operative in human cancer. Cell Reports, 2013, 3(1): 246–259. [doi: 10.1016/j.celrep.2012.12.008]

4 Rusch HP, Baumann CA. Tumor production in mice with ultraviolet radiation. The American Journal of Cancer, 1939, 35(1): 55–62. [doi: 10.1158/ajc.1939.55]

5 Reis-Filho JS. Next-generation sequencing. Breast Cancer Research, 2009, 11(3): S12. [doi: 10.1186/bcr2431]

6 Dewey FE, Grove ME, Pan CP, *et al*. Clinical interpretation and implications of whole-genome sequencing. JAMA, 2014, 311(10): 1035–1045. [doi: 10.1001/jama.2014.1717]

7 Bozic I, Antal T, Ohtsuki H, *et al*. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(43): 18545–18550. [doi: 10.1073/pnas.1010978107]

8 Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: Origins, consequences, and clinical use. Cold Spring Harbor Perspectives in Biology, 2010, 2(1): a001008. [doi: 10.1101/cshperspect.a001008]

9 Bailey MH, Tokheim C, Porta-Pardo E, *et al*. Comprehensive characterization of cancer driver genes and mutations. Cell, 2018, 173(2): 371–385. [doi: 10.1016/j.cell.2018.02.060]

10 Pittet JF, Mackersie RC, Martin TR, *et al*. Biological markers of acute lung injury: Prognostic and pathogenetic significance. American Journal of Respiratory and Critical Care Medicine, 1997, 155(4): 1187–1205. [doi: 10.1164/ajrccm.155.4.9105054]

11 Cao XR, Liu RW. General approach to blind source separation. IEEE Transactions on Signal Processing, 1996, 44(3): 562–571. [doi: 10.1109/78.489029]

12 Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788–791. [doi: 10.1038/44565]

13 Wang YX, Zhang YJ. Nonnegative matrix factorization: A comprehensive review. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(6): 1336–1353. [doi: 10.1109/TKDE.2012.51]

14 Fischer A, Illingworth CJR, Campbell PJ, *et al*. EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology, 2013, 14(4): 1–10. [doi: 10.1186/gb-2013-14-4-r39]

15 Cohen J, Cohen P, West SG, *et al*. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: Routledge, 2013. [doi: 10.4324/9780203774441]

16 Maura F, Degasperi A, Nadeu F, *et al*. A practical guide for mutational signature analysis in hematological malignancies. Nature Communications, 2019, 10(1): 2969. [doi: 10.1038/s41467-019-11037-8]

17 Koboldt DC, Zhang Q, Larson DE, *et al*. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research, 2012, 22(3): 568–576. [doi: 10.1101/gr.129684.111]

18 Saunders CT, Wong WSW, Swamy S, *et al*. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics, 2012, 28(14): 1811–1817. [doi: 10.1093/bioinformatics/bts271]

19 Larson DE, Harris CC, Chen K, *et al*. SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics, 2012, 28(3): 311–317. [doi: 10.1093/bioinformatics/btr665]

20 Lai ZW, Markovets A, Ahdesmaki M, *et al*. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Research, 2016, 44(11): e108. [doi: 10.1093/nar/gkw227]

21 Ramos AH, Lichtenstein L, Gupta M, *et al*. Oncotator: Cancer variant annotation tool. Human mutation, 2015, 36(4): E2423–E2429. [doi: 10.1002/humu.22771]

22 Pedersen BS, Quinlan AR. Cyvcf2: Fast, flexible variant analysis with Python. Bioinformatics, 2017, 33(12): 1867–1869. [doi: 10.1093/bioinformatics/btx057]

23 Lee J, Lee A, Lee JK, *et al*. Mutalisk: A Web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Research, 2018, 46(W1): W102–W108. [doi: 10.1093/nar/gky406]

24 Bamford S, Dawson E, Forbes S, *et al*. The COSMIC (Catalogue Of Somatic Mutations In Cancer) database and website. British Journal of Cancer, 2004, 91(2): 355–358. [doi: 10.1038/sj.bjc.6601894]

25 Roberts SA, Lawrence MS, Klimczak LJ, *et al*. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nature Genetics, 2013, 45(9): 970–976. [doi: 10.1038/ng.2702]