计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(10):39−45 [doi: 10.15888/j.cnki.csa.006591]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

肿瘤微卫星不稳定检测方法综述①

陈　玮[1,2], 赵　丹[1,2], 李晓东[1,2], 何小雨[1,2], 李瑞琳[1,2], 牛北方[1,2,3]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学 计算机与控制学院, 北京 100049)
[3](贵州大学 医学院, 贵阳 550025)
通讯作者: 牛北方, E-mail: bniu@sccas.cn

摘　要: 微卫星是广泛分布在真核生物基因组中的短串联重复序列. 微卫星不稳定 (Microsatellite Instability, MSI) 是指由 DNA 错配修复系统故障引起的微卫星区域重复序列插入或缺失的现象. 微卫星不稳定的检测对于肿瘤的早期诊断以及预后判断等具有重要的意义. 临床上采用 MSI-PCR 以及 MMR-IHC 的实验方法检测 MSI, 随着下一代测序技术的发展, 基于高通量测序数据的 MSI 检测方法及软件逐渐涌现. 本文将从生物学实验方法和计算方法两个角度对当前的 MSI 检测方法进行介绍并讨论分析这些方法的优势及局限.

关键词: 肿瘤; 微卫星不稳定; MSI-PCR; MMR-IHC; 统计方法; 机器学习

引用格式: 陈玮,赵丹,李晓东,何小雨,李瑞琳,牛北方.肿瘤微卫星不稳定检测方法综述.计算机系统应用,2018,27(10):39–45. http://www.c-s-a.org.cn/1003-3254/6591.html

Review on Tumor Microsatellite Instability Detection Methods

CHEN Wei[1,2], ZHAO Dan[1,2], LI Xiao-Dong[1,2], HE Xiao-Yu[1,2], LI Rui-Lin[1,2], NIU Bei-Fang[1,2,3]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)
[3](School of Medicine, Guizhou University, Guiyang 550025, China)

Abstract: Microsatellites are short repetitive sequences widespread in eukaryotic genomes. Microsatellite instability, the gain or loss of repeat units from repetitive DNA tracts caused by MisMatch Repair Deficiency (MMRD), is significant for tumor early diagnosis and prognosis. In current clinical practice, microsatellite instability detection is performed by experimental methods like MSI-PCR and MMR-IHC. With the development of next-generation sequencing technology, a number of MSI detection softwares utilizing high-throughput sequencing data have been developed. In this paper, we provide an overview on existing MSI detection methods, including experimental methods and computational methods, as well as their strengths and limits.

Key words: tumor; microsatellite instability; MSI-PCR; MMR-IHC; statistical method; machine learning

微卫星, 即短串联重复序列, 是广泛分布在真核生物基因组中的 (原核生物基因组中也有少量分布), 以 1-6bp 为一个重复单元, 重复次数不超过 60 次的 DNA 序列. 人类基因组中有数以万计的微卫星位点,

① 基金项目: 国家重点研发计划 (2016YFC0503607, 2016YFB0201901); 国家自然科学基金 (31771466); 青海省科技成果转化专项 (2016-SF-127); 中国科学院信息化专项 (XXH13504-08); 中国科学院战略性先导科技专项 (A 类) (XDA12000000); 中国科学院“百人计划”(牛北方)
Foundation item: National Key Research and Development Project of China (2016YFC0503607, 2016YFB0201901); National Natural Science Foundation of China (31771466); Commercialization of Research Findings Special Project of Qinghai Province (2016-SF-127); Informatization Special Project of Chinese Academy of Sciences (XXH13504-08); Strategic Priority Program of Chinese Academy of Sciences (Class A) (XDA12000000); “Hundred Talents Program” of Chinese Academy of Sciences (Niu Beifang)
收稿时间: 2018-03-06; 修改时间: 2018-04-03; 采用时间: 2018-04-12; csa 在线出版时间: 2018-09-28

© 中国科学院软件研究所 http://www.c-s-a.org.cn

这些微卫星位点近似均匀地分布在各个染色体上, 所有的微卫星序列约占整个基因组的 3%. 微卫星按照重复单元的大小可分为单核苷酸、二核苷酸、三核苷酸、四核苷酸、五核苷酸、六核苷酸重复; 按照重复序列的结构可分为简单重复 (由单一重复单元构成) 和复合重复 (由重复单元不同的多个重复序列构成). 与 DNA 中的其他区域相比, 微卫星区域具有较高的突变率[1]. 其高突变率的直接表现是高度的多态性[2], 即不同个体之间或正常组织与肿瘤组织之间, 微卫星位点重复单元的重复次数存在差异. 由于微卫星位点的分布广泛性及高度多态性, 微卫星常用于个体鉴定、连锁图谱的绘制以及肿瘤发生机制的研究.

微卫星不稳定 (Micro Satellite Instability, MSI), 是指微卫星位点重复单元的重复次数出现波动的现象, 即重复单元的插入与删除. 现普遍认为这种现象是由 DNA 在复制过程中出现“链滑”(strand slippage) 引起的. DNA 在复制过程中, DNA 聚合酶沿模板链滑动, 子链与模板链会发生局部分离和重新配对. 在重新配对的过程中, 子链与模版链发生错配, 就会导致一个或几个重复单元形成凸环. 一般情况下, 这种错误会被 DNA 的错配修复系统 (Mismatch Repair, MMR) 修复, 然而, 当 MMR 中的相关基因由于启动子超甲基化或基因突变等原因出现故障, DNA 复制错误无法被修复, 一些微卫星位点重复单元的重复次数发生波动, 进而发生微卫星不稳定[3]. 不同的微卫星位点稳定性不同. 微卫星重复单元的大小、重复单元的碱基组成、重复序列的结构及重复次数等都会在一定程度上影响位点的稳定性. 根据微卫星不稳定的程度, 可以将 MSI 分为 MSI-H (MicroSatellite Instability High), MSI-L (MicroSatellite Instability Low) 以及 MSS (MicroSatellite Stable).

1993 年, Aaltonen 等人首次在家族性遗传性结直肠癌 (Hereditary Non-Polyposis Colorectal Cancer, HNPCC) 中发现高频率的 MSI[4]. 微卫星不稳定在大约 15% 的结直肠癌以及 90% 的林奇综合症 (HNPCC, 又称 Lynch Syndrome) 中起决定作用[5]. 近年来的研究表明, MSI 对林奇综合症以及结直肠癌的诊断、预后以及化疗敏感性有重要的意义. 除了结直肠癌, 研究人员也相继在子宫内膜癌、卵巢癌[6]、胃癌以及乳腺癌[7]等疾病中发现 MSI. MSI 作为肿瘤遗传不稳定的敏感指标, 其检测对于肿瘤的早期诊断、预后判断、化疗敏感性判断以及高危人群的圈定等具有重要意义. 已有不少研究发现 MSI-H 的肿瘤患者相对于 MSS 的肿瘤患者有更好的预后[8,9], 同时 MSI-H 肿瘤患者对不同化疗方法的敏感性也表现出差异[10].

目前, 临床上主要采用 MSI-PCR 以及 MMR-IHC 的方法进行微卫星不稳定的检测. 然而, 近年来, 随着下一代测序技术 (Next-Generation Sequencing technology, NGS) 的飞速发展, 测序价格以超越摩尔定律的速度急速下降, 测序速度也大幅提升, 这使得方便快捷地获取测序数据成为可能. 目前已有多个通过分析测序数据来检测微卫星不稳定的软件方法.

下文将从基于生物学实验的方法和基于计算的方法两个角度来介绍现有的微卫星不稳定的检测方法.

1 基于生物学实验的方法

当前临床上主要采用聚合酶链式反应 (Polymerase Chain Reaction, PCR) 或免疫组织化学 (Immuno Histo Chemistry, IHC) 染色的方法检测患者的 MSI 状态.

MSI-PCR[11]通过对肿瘤组织和正常组织中选定的微卫星位点进行 PCR 扩增及凝胶电泳, 通过比较两组电泳结果的差异来确定 MSI 的状态. 然而人类基因组中有数以万计的微卫星位点, 不同的位点对于检测 MSI 的敏感性和准确性也各不相同. 为了标准化 MSI 的检测, NCI (National Cancer Institute) 于 1997 年推荐了 Bethesda 指南[12], 该指南推荐了两个单核苷酸位点 (BAT-25, BAT-26) 以及三个二核苷酸位点 (D2S123, D5S346 和 D17S250) 作为检测 MSI 的微卫星标记, 检测结果中有两个及以上位点出现不稳定为 MSI-H, 一个位点出现不稳定为 MSI-L, 没有位点出现不稳定为 MSS. 鉴于二核苷酸位点在对 MMR 故障的肿瘤患者的 MSI 检测中, 敏感性和准确性不及单核苷酸位点, NCI 又于 2004 年对 Bethesda 指南进行了修订[13]. 与此同时, Bacher 等人[14]通过对 266 个微卫星位点 (其中包括单核苷酸、二核苷酸、四核苷酸以及五核苷酸微卫星位点) 检测的敏感性及准确性进行评估, 提出了 Promega 分析系统, 该系统使用五个单核苷酸微卫星位点 (BAT-25, BAT-26, NR-21, NR-24 和 MONO-27) 检测 MSI, 并使用两个五核苷酸微卫星位点 (Penta C 和 Penta D) 标识样本.

与 MSI-PCR 不同, MMR IHC 通过检测 MMR 蛋白 (MLH1、MSH2、MSH6 和 PMS2) 的表达来确定

© 中国科学院软件研究所 http://www.c-s-a.org.cn

MMR 系统是否发生故障, 进而判断 MSI 的状态. 然而并不能用 MMR IHC 完全替代 MSI PCR, 因为在确定为 MSI-H 的肿瘤中, 有 5% 的肿瘤, 四种蛋白都表达, 使用 MMR IHC 无法将其识别.

2 基于计算的方法

目前, 已有多个通过分析高通量测序数据检测微卫星不稳定的方法及软件. 从模型的角度可以将这些方法分为基于一般统计模型的方法和基于机器学习模型的方法. 其中, 基于统计的方法, 首先选取一个可以反映微卫星不稳定特点的指标, 然后在一组给定的样本上 (MSI 的临床检测结果已知), 确定该指标与临床检测结果的一致性及分类阈值. 基于机器学习的方法, 主要通过特征提取、特征选择及分类器训练的方法进行 MSI 状态的预测. 不论是统计方法中的指标还是机器学习方法中的特征, 其选择的主要依据是微卫星不稳定这一现象以及其背后的产生机制. 其中, 现象, 即测序数据中表现出的微卫星位点重复单元重复次数的波动, 其本质上是碱基的插入与删除; 产生机制, 即 DNA 错配修复系统相关基因启动子超甲基化或发生突变使得这些基因无法表达, 进而影响到错配修复系统的功能. 因此, 基于计算的方法一般是通过对测序数据、超甲基化数据、突变数据以及基因表达数据进行分析, 确定 MSI 状态的.

从样本的角度可以将这些方法分为基于配对的肿瘤-正常样本的方法和仅基于肿瘤样本的方法. 第二种方法在缺乏与肿瘤样本配对的正常样本的情况下, 可以有效解决 MSI 的检测问题.

表 1 从以上两个维度对现有的方法进行了分类.

表 1 MSI 检测方法分类

	基于一般统计模型的 MSI 检测方法	基于机器学习模型的 MSI 检测方法
配对的肿瘤-正常样本	MSIsensor[15], MANTIS[16]	MOSAIC[17]
肿瘤样本	Lu, Soong [18], mSINGS[19]	MSIseq[20], MIRMMR[21]

以下将从模型的角度分类介绍各个方法.

2.1 基于一般统计模型的 MSI 检测方法

目前, 主要有以下四种基于一般统计模型的 MSI 检测方法, 这些方法均是通过对测序数据的分析, 从微卫星位点重复单元重复次数波动的角度出发, 解决这一问题的.

(1) 基于 Indel 的 MSI 检测方法[18]

MSI 中发生的重复单元的插入与删除从本质上是小片段碱基的插入与删除, 即 Indel. Lu 等人正是从这个角度出发, 将 MSI 的判定问题转化为了微卫星区域的 Indel 变化问题.

对于每个样本, 首先进行 Indel 识别, 其次对获得的 Indel 进行过滤并保留位于微卫星区域的 Indel. 通过在一组样本 (MSI 临床检测结果已知) 上对 PI、PD 以及 PI/PD 作为 MSI 判别指标进行 t 检验评估 (其中 PI 表示微卫星区域 insertion 占所有 insertion 的比例, PD 表示微卫星区域 deletion 占所有 deletion 的比例, PI/PD 为二者的比率), 选择了 PI/PD 作为样本的 MSI 判别指标. MSI-H 的样本在该指标上的取值显著低于 MSS 的样本.

Lu 等人仅提供了上述方法的工作流程并通过实际的数据验证了该方法的有效性, 并没有开发出相应的软件工具.

(2) mSINGS[19]

mSINGS 首先判断每个微卫星位点的稳定性, 进一步根据不稳定的微卫星位点的比例来判断样本的 MSI 状态. 对于每个微卫星位点, mSINGS 试图找到一个指标来量化其稳定程度, 并基于一组 MSS 样本建立各微卫星位点该指标的参考值, 对于给定样本的某个微卫星位点, 若该指标取值超出参考范围, 则认为该微卫星位点不稳定. 通过这种方式, mSINGS 解决了仅有肿瘤样本情况下 MSI 的判定问题. 具体方法如下:

1) 对于任一微卫星位点, 以其等位基因的个数作为衡量该位点是否稳定的指标, 计算一组 MSS 样本上, 该位点等位基因个数的平均值作为参考值. 具体计算方法如下:

① 仅选择在该位点测序深度大于等于 30 的 MSS 样本参与计算;

② 对每个符合条件的样本, 计算该位点等位基因的分布信息, 如表 2 所示;

③ 对每个符合条件的样本, 规范化其等位基因的支持 reads 数: 规范化的支持 reads 数=支持 reads 数/最大支持 reads 数;

④ 对每个符合条件的样本, 过滤掉规范化的支持 reads 数小于 5% 的等位基因, 以剩余的等位基因数作为该样本该位点的等位基因数;

© 中国科学院软件研究所 http://www.c-s-a.org.cn

⑤ 计算符合条件的样本该位点等位基因数的平均值 (该微卫星位点的参考值) 及方差.

2) 对于给定样本, 采用与 1) 相同的处理方式, 对比 1) 中建立的参考值, 根据 3σ 法则判断其各微卫星位点的稳定性;

3) 计算不稳定微卫星位点的比例以判定样本的 MSI 状态.

从上述 mSINGS 的方法介绍可以看出, 各微卫星位点稳定性指标的参考值是影响 mSINGS 准确性的重要因素, 而参考值的计算依赖于合理地选择一组 MSS 样本. 为了保证判别的准确率, 用于参考的 MSS 样本与待检测的样本应该具有较好的一致性, 如测序、癌种方面的一致性. 在实际使用中, 常常需要自行建立参考值.

表 2 等位基因分布信息

等位基因	[AT]0	[AT]1	[AT]2	[AT]3	[AT]4	[AT]5	[AT]6	[AT]7	[AT]8	…
支持 reads 数 (正常)	0	0	0	0	0	26	5	0	0	…
支持 reads 数 (肿瘤)	0	1	5	6	14	12	7	3	2	…

(3) MSIsensor[15]

与 mSINGS 相似, MSIsensor 也是通过分别判断每个微卫星位点的稳定性, 然后以不稳定微卫星位点的比例作为 MSI 得分. 不同的是, MSIsensor 需要基于配对的肿瘤-正常样本进行 MSI 的判定. 首先, 对于在肿瘤和正常样本中测序深度均大于等于 20 的微卫星位点, 计算其等位基因的分布信息; 其次, 通过卡方检验比较肿瘤和正常样本的相同微卫星位点的等位基因分布, 若显著不同, 则认为该微卫星位点不稳定; 最后统计不稳定位点的比例, 若该比例超过阈值, 则判定为 MSI-H, 其中, 阈值是通过该指标在一组样本上 (包括 MSI-H 和 MSS 的样本) 的累积分布确定的.

(4) MANTIS[16]

类似于 MSIsensor, MANTIS 也获得了肿瘤-正常配对样本在每个微卫星位点的等位基因分布信息; 与 MSIsensor 不同的是, 对于每个微卫星位点, MANTIS 把上述两组数据看作两个向量, 定义这两个向量的 L_1 范数为样本中该位点的稳定程度, 对所有位点的 L_1 范数求平均值即为样本的 MSI 得分. 具体方法如下:

对于每个微卫星位点,

1) 仅保留读长、测序质量符合要求的比对到该位点的 reads;

2) 分别计算配对的肿瘤-正常样本中该位点的等位基因分布;

3) 根据 3σ 法则, 过滤掉配对的肿瘤-正常样本在该位点支持 reads 不足的等位基因;

4) 经过上述处理, 仅保留在配对的肿瘤-正常样本中支持 reads 总数 (该位点的测序深度) 均超过一定阈值的微卫星位点.

5) 分别规范化肿瘤-正常样本该位点等位基因的支持 reads 数: 规范化的支持 reads 数=支持 reads 数/该位点的总支持 reads 数;

6) 根据规范化后的支持 reads 数, 计算配对的肿瘤-正常样本中该微卫星位点等位基因分布的 L_1 范数;

7) 以所有位点 L_1 范数的平均值作为样本的 MSI 得分.

MANTIS 对参与计算的数据进行了相对严格的质量控制, 如上述流程中的 1)、3) 及 4) 步骤. 由于测序过程中总会产生误差和错误, 通过质量控制, 仅使用符合要求的数据参与计算, 可以在一定程度上提高后续分析的准确性.

上述基于一般统计模型的 MSI 检测方法通过设计一个 MSI 判定指标, 在一组样本上, 使用累积分布等方式, 确定该指标的阈值, 实现对 MSI 状态的检测. MANTIS 一文从 MSI 判定的准确性及计算资源使用两个方面对 mSINGS、MSIsensor 以及 MANTIS 三种方法进行了评估, 阈值、用于分析的微卫星位点的数量以及癌种都会影响软件的准确性. 尽管在敏感度和特异度方面有细微差异, 三个软件工具均可以准确的检测样本的 MSI 状态. 然而, 不同于 mSINGS 和 MANTIS, MSIsensor 没有对等位基因分布中的支持 reads 数进行规范化以及质控, 在配对的肿瘤-正常样本测序深度不同的情况下, 可能出现假阳性的结果.

2.2 基于机器学习模型的 MSI 检测方法

目前, 基于机器学习模型的 MSI 检测方法主要有以下三种. 特征和算法是机器学习的重要组成, 以下将从这两个方面介绍各个方法. 关于每个特征的提取方式不在此赘述.

(1) MSIseq[20]

发生微卫星不稳定的样本其单核苷酸替代 (Single Nucleotide Substitution, SNS) 率以及小片段碱基的插

© 中国科学院软件研究所 http://www.c-s-a.org.cn

入与删除 (Indel) 比率都会发生变化, MSIseq 主要是从基因变异这一角度出发选取特征的. 备选特征如表 3 所示.

表 3 MSIseq 备选特征

特征	含义
T.sns	单核苷酸替代率
S.sns	微卫星区域单核苷酸替代率
T.ind	小片段碱基插入删除比率
S.ind	微卫星区域小片段碱基插入删除比率
T	T.sns + T.ind
S	S.sns + S.ind
S.sns/T.sns	
S.ind/T.ind	
S/T	
癌种	结肠癌、直肠癌、子宫内膜癌、胃癌

在这些特征的基础上, MSIseq 使用五折交叉验证分别基于 LR、决策树、随机森林、朴素贝叶斯算法训练了分类器并评估了性能, 最终选择基于决策树算法的分类器, 该分类器仅使用了 S.ind 这一个特征.

由于 MSIseq 提取的特征并不依赖于配对的肿瘤-正常样本, 因此这一方法适用于仅有肿瘤样本的情况.

(2) MOSAIC[17]

MOSAIC 是基于对每个微卫星位点稳定性的判断设计特征的. 除了与各微卫星位点稳定性相关的特征外, 还增加了通过在一组样本上对所有微卫星位点的稳定性进行分析后发现的显著不稳定的微卫星位点信息, 备选特征如表 4 所示.

表 4 MOSAIC 备选特征

特征	含义
peak_avg	肿瘤样本相对于配对的正常样本在所有微卫星位点新增的等位基因数的均值
peak_var	肿瘤样本相对于配对的正常样本在所有微卫星位点新增的等位基因数的方差
num_unstable	不稳定的微卫星位点数
prop_unstable	不稳定的微卫星位点百分比
defbsite	位于 DEFB105A/B 处的一个微卫星位点, 可以最显著地区分 MSI 和 MSS
microsatelites	MSI-H 相对与 MSS, 最显著不稳定的前 100 个微卫星位点

其中, 微卫星位点不稳定性的确定采用高敏感度的方法, 数据处理过程与 mSINGS 相同, 不同点在于微卫星位点不稳定的判定不再依据 3σ 原则, 而是对于任意微卫星位点, 若肿瘤样本相对于配对的正常样本, 在该位点出现新增的等位基因, 即认为该位点不稳定.

MOSAIC 分别基于决策树和随机森林算法训练了模型, 最终选择了基于决策树算法的分类器, 该分类器仅使用了 peak_avg 以及 defsite 两个特征 .

MOSAIC 选择的特征依赖于配对的肿瘤-正常样本, 因此仅适用于有配对样本的情况.

(3) MIRMMR[21]

与其他方法不同, MIRMMR 的特征选择主要依据 MSI 的发生机制. 使用了与 DNA 错配修复系统相关的 35 个基因的点突变率、甲基化水平以及 CADD[22] 得分作为备选特征, 基于 LR 算法构造了分类器. 相比于 MSIseq 及 MOSAIC, MIRMMR 提供了更多的建模方法, 包括 univariate、stepwise 与 penalized 三种模式. 其中 univariate 用于单变量的逻辑回归建模, 可以用于比较各特征用于 MSI 判定的准确性; stepwise 模式用于自动化的特征选择, 从备选特征中, 选择最优的特征集训练模型; penalized 模式在模型中增加了惩罚项用于防止过拟合. MIRMMR 默认使用了 penalized 模式基于 676 个样本训练了模型. MIRMMR 使用的特征不依赖于配对的肿瘤-正常样本, 因此可适用于仅有肿瘤样本的情况.

2.3 各方法的比较

针对上述提到的七种用于微卫星不稳定检测的计算方法, 从适用范围、MSI 指标、测试数据集以及软件特性等方面进行了比较, 具体如表 5 所示. 其中“—”表示无相关信息, WES (Whole Exome Sequencing) 表示全外显子组测序.

对于 MSI 的检测, 表 5 中每个方法的输出既可以是连续的 MSI 指标也可以是确定的分类. 其中基于一般统计模型的 MSI 检测方法, 可以在连续的 MSI 指标基础上, 根据阈值对样本分类; 基于机器学习模型的 MSI 检测方法, 可以预测类别也可以输出类别的概率.

从软件易用性的角度分析, MSIsensor 和 MANTIS 由于可以直接对 BAM 文件进行分析因此使用最为方便; 而 mSINGS 在对样本进行分析之前, 需要足够的 MSS 样本建立参考值, 对样本量有一定要求, 给使用带来了一定程度上的不便; 基于机器学习模型的方法, 虽然可以直接使用模型进行预测, 硬件资源使

© 中国科学院软件研究所 http://www.c-s-a.org.cn

用少, 速度快, 但是特征的提取依然是一个复杂低效的过程.

不论是基于一般统计模型的方法还是基于机器学习模型的方法, 要准确地检测 MSI 都离不开数据的支持, 测试数据集的大小和包含的癌种都会在一定程度上影响分类的准确性.

表 5 基于计算的各方法比较

	Lu, Soong	mSINGS	MSIsensor	MANTIS	MSIseq	MOSAIC	MIRMMR
适用范围	RNA-Seq	WES, 靶向测序	WES, 靶向测序	WES, 靶向测序	WES, 靶向测序	WES	—
MSI 指标	连续值	连续值	连续值	连续值	连续值	连续值	连续值
MSI 指标参考值	1	0.2	0.035	0.4	—	—	—
测试数据集	多癌种样本 324	个多癌种样本 242	个子宫内膜癌样本 458	个多癌种样本 526	个多癌种样本 617	个多癌种样本 676	个多癌种样本
软件可用性	不可用	可用	可用	可用	可用	不完整	可用
软件易用性	—	一般	易用	易用	一般	难用	一般
开发语言	—	Python	C++	Python	R	R	R
是否支持并行	—	—	支持	支持	—	—	支持

3 讨论

随着测序成本的下降和测序速度的提升, 计算方法相对于生物学实验方法的优势也越来越突出. 相比于计算方法, 通过生物学实验方法检测 MSI 有以下几个方面的不足. 首先, 需要耗费一定的时间和人力; 其次, 结果的准确性依赖于分析人员的肉眼判断; 再者, 微卫星标记和 MMR 蛋白都有其局限性. 对于微卫星标记, 实验中选择的数量有限, 存在组织 (肿瘤) 特异性[23], 无法准确地在多种肿瘤中检测 MSI 状态; 对于 MMR 蛋白, 由于 MMR 可能不是引起 MSI 的唯一原因[24], 以及 MMR 自身的复杂性, 使用 MMR 蛋白的表达来间接判断 MSI 状态也存在局限性.

计算方法利用测序数据, 从 MSI 的表现及产生机制的层面, 可以对样本的 MSI 状态作出全面的评估. 相比于生物学实验方法, 计算方法的众多优势使其可能在未来用于微卫星不稳定的临床检测. 在这个过程中, 还需要考虑以下方面的问题. 首先, 数据支持. 不论是基于一般统计模型的方法还是基于机器学习模型的方法, 要确定合适的阈值或提高分类器的准确性都需要大量数据的支持. 其次, 软件易用性. 软件要易于安装, 其使用应该在最大程度上实现自动化同时运行时间需要在可接受的范围内.

MSKCC(Memorial Sloan Kettering Cancer Center) 最近的一项研究[25]使用 MSIsensor 对 12 288 例实体癌病人的靶向测序数据进行分析, 判定 MSI 状态, 并用 MSI-PCR/MMR-IHC 进行了验证. 实验证明, 基于大规模靶向测序数据, 通过 MSIsensor 预测病人的 MSI 状态具有高的可信度. 对于 MMR 故障的样本, 相比于当前普遍使用的 MSI-PCR 方法, MSIsensor 具有更高的敏感性. 根据 2.3 节的分析, MSIsensor 使用 C++语言开发, 安装及使用十分便利, 同时支持并行计算, 运行速度快, 方便临床应用. 此项研究在一定程度上为该软件工具的临床应用提供了支持.

参考文献

1 Brinkmann B, Klintschar M, Neuhuber F, *et al*. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. American Journal of Human Genetics, 1998, 62(6): 1408–1415. [doi: 10.1086/301869]

2 丁一, 童坦君. 微卫星不稳定性的生物学意义及其应用前景. 生物科学进展, 1999, (4): 292–296.

3 Karran P. Microsatellite instability and DNA mismatch repair in human cancer. Seminars in Cancer Biology, 1996, 7(1): 15–24. [doi: 10.1006/scbi.1996.0003]

4 Aaltonen LA, Peltomaki P, Leach FS, *et al*. Clues to the pathogenesis of familial colorectal-cancer. Science, 1993, 260(5109): 812–816. [doi: 10.1126/science.8484121]

5 Pino MS, Chung DC. Microsatellite instability in the management of colorectal cancer. Expert Review of Gastroenterology & Hepatology, 2011, 5(3): 385–399.

6 King BL, Carcangiu ML, Carter D, *et al*. Microsatellite instability in ovarian neoplasms. British Journal of Cancer, 1995, 72(2): 376–382. [doi: 10.1038/bjc.1995.341]

7 Yee CJ, Roodi N, Verrier CS, *et al*. Microsatellite instability and loss of heterozygosity in breast-cancer. Cancer Research, 1994, 54(7): 1641–1644.

8 Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. Science, 1993, 260(5109):

© 中国科学院软件研究所 http://www.c-s-a.org.cn

816–819. [doi: 10.1126/science.8484122]

9 Sankila R, Aaltonen LA, Jarvinen HJ, *et al*. Better survival rates in patients with MLH1-associated hereditary colorectal cancer. Gastroenterology, 1996, 110(3): 682–687. [doi: 10.1053/gast.1996.v110.pm8608876]

10 Ribic CM, Sargent DJ, Moore MJ, *et al*. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. The New England Journal of Medicine, 2003, 349(3): 247–257. [doi: 10.1056/NEJMoa022289]

11 Berg KD, Glaser CL, Thompson RE, *et al*. Detection of microsatellite instability by fluorescence multiplex polymerase chain reaction. Journal of Molecular Diagnostics, 2000, 2(1): 20–28. [doi: 10.1016/S1525-1578(10)60611-3]

12 Boland CR, Thibodeau SN, Hamilton SR, *et al*. A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Research, 1998, 58(22): 5248–5257.

13 Umar A, Boland CR, Terdiman JP, *et al*. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. Journal of the National Cancer Institute, 2004, 96(4): 261–268. [doi: 10.1093/jnci/djh034]

14 Bacher JW, Flanagan LA, Smalley RL, *et al*. Development of a fluorescent multiplex assay for detection of MSI-high tumors. Disease Markers, 2004, 20(4-5): 237–250. [doi: 10.1155/2004/136734]

15 Niu BF, Ye K, Zhang QY, *et al*. MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics, 2014, 30(7): 1015–1016. [doi: 10.1093/bioinformatics/btt755]

16 Kautto EA, Bonneville R, Miya J, *et al*. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget, 2017, 8(5): 7452–7463.

17 Hanse RJ, Pritchard CC, Shendure J, *et al*. Classification and characterization of microsatellite instability across 18 cancer types. Nature Medicine, 2016, 22(11): 1342–1350. [doi: 10.1038/nm.4191]

18 Lu Y, Soong TD, Elemento O. A novel approach for characterizing microsatellite instability in cancer cells. PLoS One, 2013, 8(5): e63056. [doi: 10.1371/journal.pone.0063056]

19 Salipante SJ, Scroggins SM, Hampel HL, *et al*. Microsatellite instability detection by next generation sequencing. Clinical Chemistry, 2014, 60(9): 1192–1199. [doi: 10.1373/clinchem.2014.223677]

20 Huang MN, McPherson JR, Cutcutache I, *et al*. MSIseq: Software for assessing microsatellite instability from catalogs of somatic mutations. Scientific Reports, 2015, 5: 13321. [doi: 10.1038/srep13321]

21 Foltz SM, Liang WW, Xie MC, *et al*. MIRMMR: Binary classification of microsatellite instability using methylation and mutations. Bioinformatics, 2017, 33(23): 3799–3801. [doi: 10.1093/bioinformatics/btx507]

22 Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics, 2014, 46(3): 310–315. [doi: 10.1038/ng.2892]

23 Faulkner RD, Seedhouse CH, Das-Gupta EP, *et al*. BAT-25 and BAT-26, two mononucleotide microsatellites, are not sensitive markers of microsatellite instability in acute myeloid leukaemia. British Journal of Haematology, 2004, 124(2): 160–165. [doi: 10.1046/j.1365-2141.2003.04750.x]

24 Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology, 2010, 138(6): 2073–2087.e3.

25 Middha S, Zhang LY, Nafa K, *et al*. Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. Journal of Clinical Oncology Precision Oncology, 2017. [doi: 10.1200/PO.17.00084]

© 中国科学院软件研究所 http://www.c-s-a.org.cn