癌症研究中的计算思维

作者：刘立洋 学号：1400012120 院系：生命科学学院

【摘要】癌症研究中，新的治疗手段不断出现，新型药物层出不穷，但是生物体的复杂程度难以预计，癌细胞的适应能力也不断挑战科学家的水平，为了能够了解癌细胞的发育与进化过程，科学家在不断改进治疗手段的同时，也开始对癌症发生和发育进行数学建模并利用计算机对癌症的演化进行分析和预测，希望通过这种方式能够帮助了解癌症的发育和演化的过程，从而帮助癌症的治疗。这篇文章将会介绍利用计算思维去研究癌症的思路和一些具体的数学模型。

【关键词】癌症发生 进化 数学建模 计算生物学

现在已知人类最早的癌症记录是在 4500 年前，但是直至今日，虽然人类的医疗技术水平已经发展到了一个非常发达的程度，但是对待癌症，却依旧没有特别好的解决办法。现在的治疗方式除了手术切除外，就是通过药物进行治疗或放射治疗。手术切除是最彻底的一种手段，但是很多癌症并不能用手术来切除，比如白血病，比如黑色素瘤。对于不宜用手术切除的癌症，最好的办法就是药物治疗，但是在治疗过程中，人们逐渐发现，癌细胞对抗癌药物就像细菌对抗生素一样，少数癌细胞会在不断的繁殖中产生耐药性并将这种耐药性传递下去，最终导致所有存在的癌细胞都具有耐药性，因此很多病人虽然在治疗初期获得了很好的疗效，但最终却不得不抱憾而终。总的说来，我们对癌症发生中的很多细节都不甚了解，哪些细胞易产生癌症？这些癌细胞又是如何发育演化最终成为恶性肿瘤的？这些问题都亟待解决。对此，早在上世纪七十年代，就有科学家提出“我们应该集中精力弄清楚癌细胞在达到最后阶段前的演化过程。”

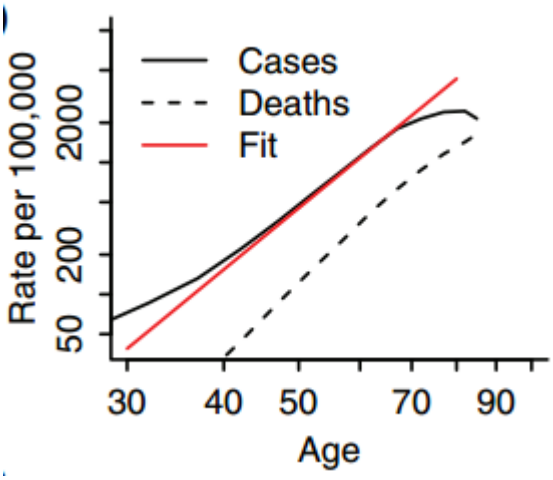
面对药物治疗中的种种无奈，科学家也开始逐步将研究重点转移到癌症发生和演化的研究上。而癌细胞和体内其他细胞不同，它们在癌变的过程中会发生各种各样的基因突变包括染色体片段的大段断裂、丢失和加倍，遗传背景非常复杂，因此也就产生了大量的异质的数据。为了能够对癌细胞演化进行有效的分析，就需要建立适当的数学模型和算法并利用计算机对巨量数据进行运算。

癌症数学建模和计算的基础数据主要有两类，一类是癌细胞的宏观表型，比如存活率等等，通过对癌细胞这些宏观表型的测量和计算，可以预测肿瘤的发生速度和生长状况；一类是海量的 DNA、RNA 测序结果，通过对比不同癌症病人、同一癌症病人的癌细胞和正常体细胞的基因，我们可以找出突变的基因位点，分析哪些位点的突变是致病突变，不同癌细胞之

间的亲缘关系和系统发生等等。对于这些基础数据，有很多方法进行分析，接下来我会介绍几种典型的方法。

一、 Multistage Theory

Multistage Theory 是由 Nordling 在 1953 年提出的，在大量的观察和统计后，他提出，癌症的发生是一个随着时间逐步发生的多步骤的过程。如下图



具体来说，他认为这是一个六个步骤的累积过程。为此，他提出了最早的预测癌症发生率数学模型：$I(t) \propto u_1 \cdots u_k t^{k-1}$ 其中 k 是经历的步骤数，u_i 是从 i-1 到 i 阶段的转化率，t 是时间。

此模型是癌症研究领域的一个突破，因为它最早对癌症的发生机理进行了预测，并且在之后的分子生物学研究中被得到证实。我觉得这是一个数学模型，通过对具体发生过程的预测，从而得到比较精准的结果，这个应该也可以算一种算法的应用。当然，这个模型也是有很多缺点的，但鉴于篇幅有限，不进行讨论。

二、 Moran process

这个过程是一个描述不同类型细胞博弈的一个算法，它有一个前提是细胞总数 N 不变。具体来讲，它将一团组织中的细胞分为两种，如癌细胞和正常细胞。并给予细胞不同的适应性 f1（正常细胞）和 f2（癌细胞）（fitness）。用 X（t）来记录在 t 时刻癌细胞的数量。为了保证细胞总数的不变，它假设每当一个细胞被随机选择进行一轮繁殖时，就会有另一个细胞被选择死亡。这样，在这样一个模型中，当随机过程进行完成后，结果是一定的，都是随 f 的变化而变化。在中性进化中，f1 和 f2 是相等的，那么癌细胞想达到占领全部数量 N 的概率就为：X/N。如果癌细胞有在进化上有选择优势（f2>f1），那么癌细胞最终消灭正常细胞的概率就为$\rho x = \frac{1-(\frac{f2}{f1})\text{^}-X}{1-(\frac{f2}{f1})\text{^}-N}$

三、Wright–Fisher process

这个模型和上面的一个模型非常相似，只是它规定代与代之间没有世代重合，计算机的运行结果证明这个用来预测肿瘤组织的算法比上一个算法更加的可靠和高效。

四、Phylogenetic Tree Reconstruction from SNVs

研究肿瘤细胞的系统发生树有利于对区分不同的肿瘤细胞和他们的发育阶段，便于对症下药。但是在单个肿瘤中往往有多个肿瘤细胞系。DNA 的测序结果又只能提供整个肿瘤组织块的信息，为了能够区别不同的细胞系确定肿瘤组织的组成，科学家们相继提出多种模型和算法，但是由于篇幅有限，我没有机会去展示这些算法，可以推荐一篇 paper，它提出了一个确定异质性肿瘤中不同细胞系的算法：TrAp: a tree approach for fingerprinting subclonal tumor composition。在这个模型中，为了使计算简化并排除一些特殊情况，作者遵循了领域内的两条假设：（1）没有突变会发生两次（2）没有突变会丢失。我觉得这是两个很好的限制条件，对比于我们平常写的代码，就像是我们给程序设置的一些不可违背的条件，可以帮助我们简化算法，同时也不失有效性。

可以看到，在研究和预测癌发生和癌细胞演化中，计算方法起到了巨大的作用。而且可以看到，每一种算法都有它自身的缺点和优点，每一种计算方法都希望能够高效准确的得到运算结果，因此都会有自己的简化步骤，但是这些简化步骤同时也给这些模型和算法留下了漏洞，所以需要进一步的改进或其他算法的补充。

【参考文献】

1. Beerenwinkel N, Schwarz R F, Gerstung M, et al. Cancer Evolution: Mathematical Models and Computational Inference[J]. Systematic Biology, 2015, 64(1):e1-e25.
2. Strino F, Parisi F, Micsinai M, et al. TrAp: a tree approach for fingerprinting subclonal tumor composition.[J]. Nucleic Acids Research, 2013, 41(17):e165.
3. Moran P.A.P. 1958. Random processes in genetics. Math. Proc. Cambridge. 54:60–71