arXiv:2305.01475v1 [q-bio.GN] 1 May 2023

用于生成具有所需基因组学特征的合成 DNA 序列的癌症启发式基因组映射模型

Teddy Lazebnik[1*] 和 Liron Simon-Keren[2]

[1]伦敦大学学院癌症研究所癌症生物学系，英国伦敦[2]特拉维夫大学机械工程学院，以色列[*]通讯作者：t.lazebnik@ucl.ac.uk

摘要

基因组数据在现代医学中至关重要，在诊断和治疗方面具有巨大潜力。得益于技术进步，目前已有数百万健康和患病个体的基因组完成测序；然而，获取适用于特定研究、尤其是验证研究的最合适数据，在数据规模和获取途径方面仍面临挑战。因此，*in silico* 基因组学序列生成器被提出作为一种可能的解决方案。然而，现有生成器主要利用浅层（随机）关联生成质量较差的数据，而这些关联是在训练数据中以有限的计算复杂度检测到的。这意味着它们没有考虑最初导致观测到这些关联的恰当生物学关系和约束。为解决这一问题，我们提出了癌症启发式基因组映射模型（CGMM），将遗传算法（GA）与深度学习（DL）方法相结合，以应对这一挑战。CGMM 模拟产生遗传变异和突变的过程，将易于获取的对照基因组转化为具有所需表型的基因组。我们证明，CGMM 能够生成具有特定表型（如祖源和癌症）的合成基因组；基于无监督聚类，这些合成基因组与具有相同表型的真实基因组无法区分。结果表明，在两项不同任务中，CGMM 的表现优于四种当前最先进的基因组学生成器，这说明 CGMM 适用于基因组医学中的广泛用途，尤其是亟需的验证研究。

关键词：i n silico 序列生成；生物标志物检测；验证；生物信息学深度学习模型。

1 引言

遗传数据在医学新时代发挥着核心作用，得益于基因测序技术的进步，它在诊断和治疗方面都展现出巨大潜力 [1–3]。然而，由于可用数据量不足 [4]，当前的医学遗传学实践很少能够开展可靠的基因组学分析。

人们已开发出多种计算方法来分析基因组学数据集，以探索种群中与表型特征和疾病相关的生殖系和体细胞遗传变异[5–7]。尽管科学界、资金支持机构、企业和公众都认识到，免费开放获取基因组学数据对科学研究和医学进步至关重要 [8, 9], ，但在实践中，基因组学数据及相关元数据的共享仍然不足，或流程过于复杂 [10]。多个项目尝试应对这一挑战。例如，个人基因组计划（PGP） [11] 以开放获取的方式提供其全部数据。癌症基因组图谱（TCGA） [12]和国际癌症基因组联盟（ICGC） [13] 则将不同部分的数据分为开放获取和受控访问两个层级。此外，数据多样性不足（例如罕见疾病、非欧洲血统人群和少数族裔代表性不足）阻碍了许多表型特异性和种群特异性模型的开发，而这些模型对于尽可能实现基因组医学的公平普惠至关重要。

为解决这些局限性，目前已提出三种主要方案。第一种是数据模拟器，它们通过计算机模拟来模拟基因型和表型数据 [14–17]。虽然这些工具解决了数据匮乏的问题，但无法充分反映数据的复杂性。第二种是数据生成器，它们生成同类型的

数据，并采用更先进的统计方法提取基因组特征，但生成的结果大多只是输入数据的轻微变体 [18–20]。这两种方法都缺乏数学形式化，限制了它们所能处理的复杂程度。这主要是因为它们考虑的参数空间较小，且这些模型使用的可行操作种类有限。

另一类基因组生成方法采用机器学习（ML）方法提取复杂模式，随后利用这些模式生成符合其特征的基因组序列 [21–23]。这些方法能够从数据中学习极其复杂的模式和特征，但往往难以应用，因为它们需要大量输入数据才能正常运行，而这通常无法获得。

为了解决这些问题，我们开发了 CGMM，这是一种新颖的数学框架，即使输入数据稀缺，也能生成合成基因组序列。与旨在捕捉单个数据集中特征的其他模型不同，我们的方法着重捕捉变异和变异如何将对照基因组转化为病例基因组的进化与机制动力学。由此，我们的癌症启发式基因组映射模型（CGMM）能够利用大量基因组作为对照集，并学习将其映射到一个较小的、具有目标表型的病例基因组集。随后，该模型能够将学到的映射应用于独立的对照基因组，将其转化为逼真的病例基因组。

具体而言，CGMM 基于一种遗传算法构建，该算法优化将对照基因组样本映射到病例基因组样本的过程。该遗传算法采用一种适应度函数，其设计受到推动癌症进展的进化机制启发，包括突变、选择与适应。该适应度函数综合考虑基因组数据的多种特征，包括序列比对、质量评分和结构变异，以准确映射合成读段。随后，CGMM 引入机器学习组件，利用循环神经网络学习并预测最优映射突变步骤。该神经网络使用大规模模拟突变模式数据集进行训练，以学习基因组特征与最优映射参数之间的关系。

我们在两个数据集上评估了 CGMM，结果表明，它能够通过生成与病例集相关的合成基因组，可靠地将目标样本数量扩增至原来的数倍。我们将生成的合成样本与四种当前最先进的基因组生成模型所生成的样本进行了比较，结果表明 CGMM 的表现优于这些模型。

本文其余部分安排如下：第 2 节概述用于 计算机模拟DNA 生成的最新计算与数学模型，包括启发 CGMM 构建的相关信息。第 3 节正式介绍 CGMM。第 5 节介绍两项 计算机模拟实验，用于评估 CGMM 相较于其他基因组生成模型的性能。最后，第 6 节总结研究结果，并提出未来可能的研究方向。

2 相关工作

本节回顾用于生成合成 DNA 序列的最新方法、癌细胞的自然进化模式，以及遗传算法在生物信息学中的优势（CGMM 模仿并实现了这些优势）。

2.1 计算机模拟DNA 序列的生成

基于各种假设生成计算机模拟DNA 序列，源于对类似病原体和样本模式的既有发现。具体而言，通过处理可用数据或请用户提供数据，定义单核苷酸多态性（SNP）的分布，然后从中随机抽样以生成新序列 [24, 25]。此外，多个模拟器假设 SNP 之间存在独立的线性联系 [16,26]。例如， [27] 提出了数量遗传学中的多性状、多基因座表型模拟，该方法基于数学框架中的模型线性混合，并考虑四个组成部分：遗传变异效应、无穷小遗传效应、非遗传协变量效应和观测噪声效应，这些效应体现了 N 样本和 p 性状的特征。这种方法使模拟器能够捕捉真实的协变量结构。然而，该模型受限于其线性数学本质。

众所周知，这些假设在实践中并不成立，因为从生物学角度来看，单核苷酸多态性并非在基因组中随机抽样，而且它们之间不必然存在线性且相互独立的关系 [28–30]。例如， [31] 表明，在癌症进展过程中，结构性和数目性染色体异常以及基因突变会非线性地累积。因此，若干模型放宽了这些假设，以便更准确地表示基因组学动态 [32–34]。

[35] 提出了一种上位性模拟流程，可生成二分、分类和定量表型；该流程能够模拟单核苷酸多态性分布中的非线性交互，同时考虑观察偏倚。该模拟由四个依次进行的计算过程构成：生成基因型语料库、对单核苷酸多态性集合进行包含处理、将上位性模型应用于抽样得到的单核苷酸多态性集合，以及对个体集合进行次级抽样。该模型的一个局限是，用户必须明确指定希望模拟的单核苷酸多态性集合。

[36] 提出了一种模拟器，旨在平衡基因型数据的真实度与计算速度。具体而言，该模拟器通过从全基因组或用户指定的基因组区域中，按照正态分布或几何分布随机抽取遗传标记作为定量性状核苷酸（QTN）。所提出的模型允许根据先前在基因组不同区域分配的 QTN，动态调整 QTN。这种动态约束分配能够实现复杂的非线性动态，但需要在设计阶段预先了解这些动态，而不是从输入数据中提取。

[37] 开发了一种模拟器，可根据用户提供的种群基因型数据模拟全基因组关联研究表型数据，并考虑遗传力、显性、群体分层以及单核苷酸多态性之间的上位性相互作用等多种生物学特性。作者将单核苷酸多态性之间的相互作用拆分为对单核苷酸多态性一维分布的因果抽样；随后，流程通过计算单核苷酸多态性之间的相互作用来丰富这些动态。之后，模拟器在种群中引入表型遗传力和群体分层，以获得病例与对照表型基因组。

2.2 癌细胞中的自然突变过程

癌症是一种复杂疾病，由细胞中遗传和表观遗传改变的累积引起，这些改变会驱动不受控增殖和侵袭 [38]。健康细胞演化为癌细胞是一个多步骤过程，涉及调控细胞周期进程、DNA损伤修复和细胞死亡等关键细胞功能的基因发生各种突变 [39, 40]。癌细胞一旦获得初始突变，就会发生克隆扩增并进一步累积其他突变，进而出现具有不同表型和遗传谱的亚克隆 [41]。因此，从癌细胞种群的角度来看，如果不加以治疗，突变的数量会不断增加。在突变过程中，演变为癌细胞的健康细胞，其突变受到一定限制，因为免疫系统能够检测到不需要的突变，并将其清除出体外 [42, 43]。同样，在短时间内以足够高的速率发生突变的细胞，通常也会被免疫系统检测到并清除出体外 [44]。总的来说，癌细胞在这一过程中会提高突变率，但每一步突变都受到一定限制。在我们的模型中，我们将这一进化方法融入遗传算法框架，并以加速有界变异算子的形式加以实现。

2.3 生物信息学中的遗传算法

遗传算法（GA）是一种受进化论启发的优化方法 [45]。具体而言，GA通过自然选择模拟“进化”过程，适应度函数得分较高的解（也称为“基因”）被认为具有更强的适应性。因此，这些基因将其信息传递给下一代的概率更高。从一代到另一代的进化跃迁通过三种随机过程实现：变异 [46],交叉 [47], 和可行性测试 [48]。

遗传算法已应用于工程 [49], 医学 [50], 和经济学 [51]等多个领域。 [48] 研究了基于机器序列及其执行操作的工艺规划优化问题。作者首先使用遗传算法获得可行工艺，随后从可行工艺集合中找出最优工艺。 [47] 探讨并分析了遗传算法在工艺路线排序和严格性方面应对各种约束的应用。作者重构了遗传算法，包括建立编码策略、评价算子和适应度函数，结果表明，新遗传算法能够满足排序工作的要求以及严格性要求。

在基因组学数据领域，人们已使用遗传算法应对多项生物信息学挑战。例如， [52] 研究了不同的遗传算法算子在将脱氧核糖核酸（DNA）序列片段组装成与母序列相对应的共识序列方面的效果；这些片段来自一个序列未知的亲本克隆，研究结果颇具前景。此外， [53]分析了多种癌症类型的基因表达数据，包括卵巢癌、前列腺癌和肺癌。作者提出了一种用于基因表达数据分析的集成式基因搜索算法，该算法基于遗传算法方法，并采用基于相关性的启发式方法进行数据

预处理。该算法将44个DNA样本准确分类为三种癌症类型，准确率达到89%。总体而言，遗传算法在生物信息学研究中展现出良好的应用前景 [54–57],，尤其是在基因组学相关任务中 [58, 59]。因此，遗传算法是生成合成DNA序列的一种很有前景的计算方法。

3 模型定义

与许多其他数据驱动模型一样，CGMM 的使用分为两个阶段：训练和推理。在训练阶段，CGMM 会解决若干优化任务，以获得构成 CGMM 的内部机器学习组件的权重。随后，在推理阶段，CGMM 接收此前未见过的新样本并作出预测。图 1 示意展示了 CGMM，概述了训练和推理过程中的主要计算步骤。

在训练阶段，CGMM 会执行四个主要步骤。第一步是获取所需输入：采用变异调用格式（VCF）文件格式的对照集和病例集。对照集和病例集中的样本会按照用户定义的比例划分为两部分，一部分用于训练阶段，另一部分留作推理阶段。此外，用户可以提供与病例组相关的单核苷酸多态性，供 CGMM 训练时使用。随后，CGMM 会根据用户自行选择的某种配对方法（II），从训练数据集中生成对照样本与病例样本配对。概括而言，配对方法主要有三种：1）暴力枚举方法，即将对照集中的每个样本与病例集中的所有样本配对。此方法生成的配对数量最多，因此训练数据集也最大。然而，当配对数量达到一定程度后，每新增一组配对带来的边际贡献会变得微不足道，从而导致不必要的计算。2）以个体为中心的方法，即从同一个体中选取对照样本和病例样本进行配对。这样可以保留最终合成的 DNA 序列中的个人特征，该序列包含病例突变。3）缩减的暴力枚举方法，即使用暴力枚举方法生成的配对组中的任意子集。具体而言，由于最终目标是学习从对照集到病例集的配对关系，因此可以为每个对照病例均匀选取 $z \in \mathbb{N}$ 个随机病例样本。尽管此方法给模型引入了另一层随机性，但只要使用得当，就能在保留暴力枚举方法优势的同时消除其缺点。因此，该方法是 CGMM 的默认方法。

获得对照样本和病例样本后，CGMM 开始第二步训练，使用反向生物过程遗传算法 (RBGA)生成一组可能的突变过程，使对照 样本转变为病例 样本。形式上，RBGA 是一种经过多项修改的遗传算法（GA）。总体而言，GA 遵循相同的计算框架，包含四种主要算子：停止条件、选择、交叉和变异。首先，GA 会随机生成一个候选解种群。然后，停止条件算子会在满足停止条件时指示 GA 停止；否则，GA 将迭代执行其余三种算子。选择算子用于从种群中选出个体，以供下一轮迭代使用。该算子旨在通过适应度函数进行随机选择，生成一个新的解种群，使其平均而言更适合当前任务。交叉算子通过组合两个个体的信息来生成新个体，模拟繁殖和生物学交叉。最后，变异算子会对种群中的个体引入随机改变。这些算子各有多种实现方式，直接决定了 GA 的具体形式。对于 RBGA，我们将适应度函数设为种群个体与病例 样本之间的 MASH度量 [60] 。据此，停止条件算子设为$\frac{1}{N}\sum_{i\in N} MASH(p_i,\ C) \leq \xi$，其中 $N \in \mathbb{N}$是 GA 的种群规模，以 P表示； $p_i \in P$是当前 GA 种群中的一个个体； C 是病例 样本； $\xi \in \mathbb{R}^+$ 是用户设定的阈值。选择算子（也称为下一代 算子）采用“贵族锦标赛” [47]。我们选择这一选择算子，是因为多项研究表明，它的表现优于其他算子，并且能够应对复杂的多维优化任务，例如我们的任务 [61]。对于交叉算子，我们从 GA 种群中随机（服从均匀分布）选取个体配对，并随机选取一个枢轴点，从而为 GA 种群生成两个新个体。通过按照随机枢轴点 [46]交叉原始两个个体各自的两部分，这两个新个体将取代原来的个体。最后，我们提出一种受癌症变异方式启发的新型突变过程。与癌细胞（代表病例 样本）由健康细胞（代表对照 样本）演化而来的方式类似，CGMM 会在对照 DNA 序列中删除、添加或编辑 SNP，同时保持其与原始序列的相似性。为此，我们采用自适应梯度下降算法 [62], 中的递减学习率方法，同时限制单次变异步骤所能造成的

总距离。具体而言，设 $\mu \in \mathbb{R}^+$ 和 $\sigma \in \mathbb{R}^+$ 分别为随机变量 $\zeta \sim N$的正态分布的均值和标准差，记作μ，σ^2。此外，定义折扣因子 $\lambda \in (0, 1)$。加速有界变异算子的运作方式如下：对于 GA 的第 i_{th}代，我们随机选取一个变异次数 ζ_i，其服从 $N(\mu \cdot \lambda^i, \sigma^2 \cdot \lambda^i)$分布。随后，我们对该个体施加 ζ_i 次改变，这些改变通过随机选取 SNP 来实现。如果变异后个体与变异前样本之间的 MASH度量距离大于预先设定的阈值 $\beta \in \mathbb{R}^+$，则该变异个体被视为不可行，并重复此过程。此外，在此阶段，如果用户提供了与病例组相关的 SNP，就会计算这些 SNP 在病例组中的分布，并将其用作变异算子的分布函数（取代均匀分布函数）。

在预先设定数量的样本对（或未另行设置时的全部样本对）生成突变过程后，CGMM 将启动训练阶段的第三步，在这一步中训练一个带有自注意力层的自编码器（AE）模型。具体而言，AE 接收一个由数字构成的列表作为突变过程，其中数字表示基因组数组中的索引以及 12 种可能的单核苷酸多态性变异之一。AE 使用 AutoKeras 自动化深度学习库生成，以便为插入的特定输入找到近乎最优的 AE 架构 [63] 。为此，AutoKeras 框架通过对编码层的大小 [64]进行贝叶斯优化，力求最小化输入样本与重构样本之间的均方误差。获得 AE 模型后，RBGA 生成的突变过程中每一步都会转换为一个编码向量。由此得到一组维度相同（即 AE 编码层的维度）的向量，用于描述一个可能的突变过程，该过程将特定的一对对照 和病例 样本关联起来。

在训练阶段的第四步，也是最后一步中，CGMM 使用循环神经网络（RNN）架构和一个长短期记忆层（LSTM），以预测 对照 样本与病例 样本对之间可能发生的突变过程的下一步。CGMM 的这一 RNN 组件称为下一步变异预测器（NMP）。NMP 作为时间序列预测器工作，其原理类似于大型语言模型学习预测句子中的下一个词 [65]。具体而言，它采用了 [66], 提出的模型架构，因为该架构在下一个词预测任务中表现出良好的准确性，同时模型规模相对较小，因此计算成本也较低。与 [66], 的研究类似，NMP 每次到达编码表示向量末尾时，都会接收一个特殊标记，以指示突变过程已经结束。这会告知 NMP 特定突变过程已结束，并防止它无限期地尝试预测该过程的下一步。

最后， 训练完成的 NMP 和自编码器模型会被保存，以供推理阶段后续使用 .

在推理阶段，CGMM 会执行三个主要步骤，此时输入的是为推理预留的 $\alpha \in \mathbb{N}$对照样本。第一步，CGMM 使用训练期间生成的自编码器的编码器部分，将样本编码为编码向量。第二步，将编码向量输入 NMP 模型，预测它们各自转变为病例 样本的突变过程。最后，将预测结果输入自编码器的解码器部分进行解码，将其转换回 VCF 格式序列。由此得到一组 α 合成病例 样本。

4 实验设计

我们在三项任务中测试了 CGMM，以展示其可靠扩充现有病例集 DNA 序列的能力。为此，我们从公共数据库收集基因组序列，并将其划分为需要用合成数据扩充的目标病例集，以及用于此任务的对照集 DNA 序列。具体而言，在前两项实验（实验1和实验2）中，我们使用了来自个人基因组学项目（PGP）的 300 个样本，分别来自三个国家：英国（120 个样本）、加拿大（110 个样本）和美国（70 个样本）。我们将这些样本手动划分为 200 个欧洲血统占主导的 DNA 序列，作为对照集；以及 100 个非洲血统占主导的 DNA 序列，作为病例集。为确保标注的祖源有效，我们核实了所有样本的标签既符合基于主成分分析的祖源分类，也符合受检个体自我报告的祖源 [67]。在最后一项实验（实验3）中，我们使用了来自个人基因组学项目、对应英国健康个体的 90 个对照样本，以及从基因组数据共享库（GDC）获取的 30 个病例 样本，这些样本对应患有皮肤黑色素瘤的个体。

为展示将对照样本转换为病例 样本的转换率与成功转换的对照样本绝对数量之间的平衡，实验 2 采用了 50%-50% 的训练集与测试集比例，而实验 1 和实验 3 则采用了 70%-30% 的比例。各阶段最终的样本数量见表 1。

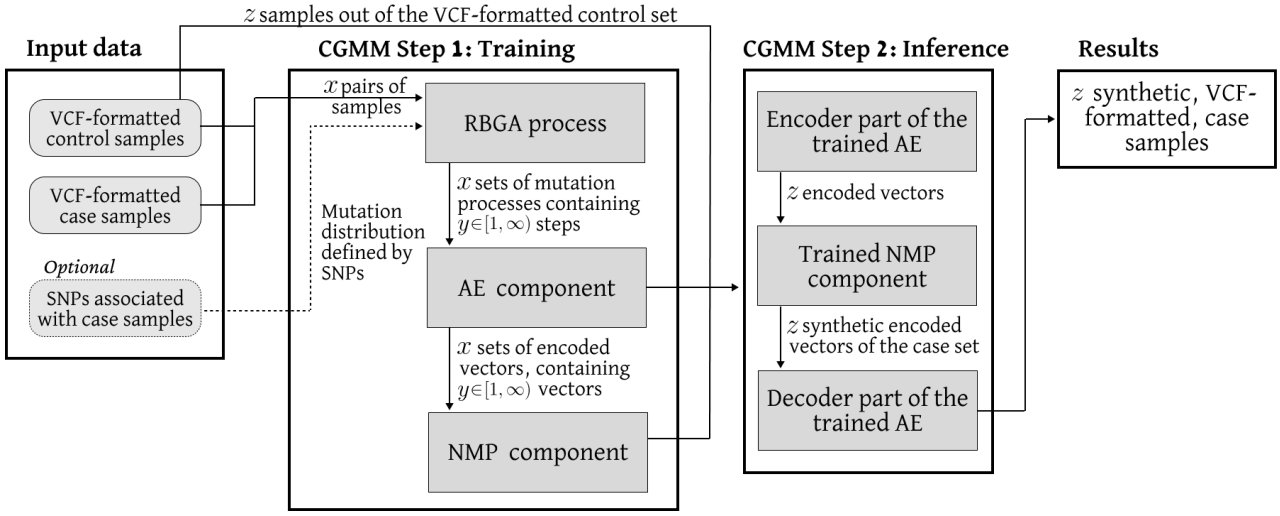


图1：模型训练阶段和推理阶段的示意图。首先，将输入数据划分出来，留出指定部分的对照样本（z 个样本）用于推理。其余输入数据进入训练阶段，在此阶段：1）估算每对对照样本和病例样本之间的演化过程（RBGA过程）。2）将该过程中的每个步骤编码到更小的潜在空间中（AE组件）。3）训练预测模型，以预测演化过程中的下一个编码步骤（NMP组件）。在推理阶段，使用 z 对照样本生成演化过程第一步的编码向量（编码器部分）。随后，经过训练的NMP组件预测一系列变异步骤，直至达到所需的基因组特征。最后，将代表编码后的合成病例样本的最后一个预测步骤解码为VCF 格式（解码器部分）。

	实验1		**实验2**		**实验3**	
	控制	*Case*	控制	*Case*	对照组	*Case*
训练	140	70	100	50	60	20
Test	60	30	100	50	30	10

表1：训练和推理阶段使用的对照组 样本和病例组 样本数量。推理时，CGMM 运行只需要对照组 样本，而所列病例组 样本则用于评估结果的可靠性。

数据集划分完成后，使用附录表 3 中列出的超参数训练模型。

为证明 CGMM 能够学习与每个病例组 相关的正确关系和基因组学特征，我们将每项实验重复两次：第一次不整合任何知识，第二次则整合外部知识，即 ForenSeqTM DNA [68] 中的祖源相关单核苷酸多态性（实验1-2），或 [69]中的皮肤黑色素瘤相关单核苷酸多态性。

为评估 CGMM 的性能，我们采用 CLINK 算法 [70],进行层次聚类，并使用 MASH 度量 [60]计算样本间的距离。尽管还有许多其他度量方法，例如詹森-香农散度（JSD） [71] 和 Kullback-Leibler（KL）散度 [72], ，我们仍决定使用 MASH，因为它在生物信息学领域广受欢迎，且已得到充分验证。随后，我们报告转换率，即推理过程中使用的全部对照组 样本中被分类为病例组 样本的对照组 样本所占的百分比。我们将 CGMM 在各项实验中的结果与其他四种模型的结果进行比较，这四种模型被视为最先进的基因组生成模型 [27, 36, 37, 73]。

5 结果

表2展示了五种受测模型将对照 样本转换为可靠病例样本的转换率。正如预期，结果表明，训练样本越多，模型就越容易

检测到与每组病例相关的基因组特征。因此，实验1中所有模型的转换率最高（不考虑包含单核苷酸多态性的情况）。然而，一旦整合了单核苷酸多态性知识，训练样本数量的影响便会显著减弱。此外，在所有实验中，CGMM受添加单核苷酸多态性知识的影响最小，在实验1-3中的转换率仅提高了15.6% $\pm$ 10.2%，而其余模型的转换率提高了36% $\pm$ 2.8%。

测试配置	CGMM	G2P [36]	Zhou 等人 [73]	PhenotypeSimulator [27]	GEPSi [37]
E_1	78.3%	38.3%	31.6%	40.0%	56.6%
E_1 含单核苷酸多态性	85.0%	71.6%	68.3%	83.3%	81.6%
E_2	73.3%	31.6%	26.6%	38.3%	41.6%
E_2 含单核苷酸多态性	86.6%	68.3%	73.3%	81.6%	78.3%
E_3	46.6%	23.3%	16.6%	36.6%	30.0%
E_3 含单核苷酸多态性	73.3%	53.3%	46.6%	66.6%	70.0%

表2：CGMM 与四种最先进模型将对照样本转换为病例样本所实现的转换率比较。

图2展示了实验1-2中最终病例DNA 序列数据集的样本数量增加情况。尽管两项实验均以相同的样本基线数量（蓝色）开始，且实验2对所有模型都产生了较低的转换率（见表2），但模型在实验2中取得了比实验1更优的结果。换言之，所有模型最终都生成了更多可靠的合成DNA序列（包括具有单核苷酸多态性知识（黄色）或不具有该知识（红色）的序列）。

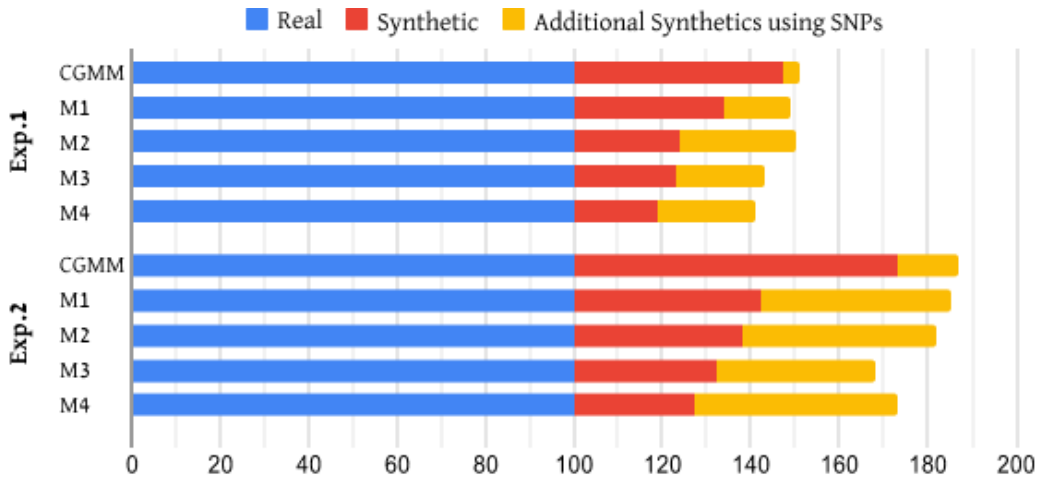


图2：病例集DNA序列的最终样本数量，包括原始真实样本（蓝色）、在不使用任何单核苷酸多态性知识的情况下生成的合成样本（红色），以及通过整合知识所实现的额外合成样本数量（黄色）。CGMM 的结果与 GEPSi（M1）、PhenotypeSimulator（M2）、G2P（M3）和 Zhou 等人（M4）所取得的结果进行了比较。

6 讨论与结论

在本研究中，我们提出了一种新型癌症启发式基因组映射模型（CGMM），它结合了遗传算法、独特的类癌症突变过程、深度学习自编码器和下一词预测模型，用于捕捉两个基因组样本之间与突变过程相关的非线性动态。借助 CGMM 模型，我们能够生成合成基因组，同时保留输入样本的个人特征。为此，模型学习样本对之间可能发生的突变演化过程，并通过模拟癌细胞演化的逻辑假设对每个演化步骤加以限制。因此，CGMM 能够接收一个输入基因组并对其进行变异，直至达到某项期望属性。因此，

我们的方法在计算方式上不同于当前的合成基因组学模型：后者试图根据具有期望属性（例如祖源或病原体）的基因组数据集，捕捉基因组层面的独特统计特性。与这些模型不同，我们旨在捕捉对照样本数据集与具有我们所需期望属性的病例 样本数据集之间的统计突变动态，而前者更容易获取。这样，CGMM 就能利用大量可获取的基因组样本，可靠地扩充另一个规模较小的基因组样本数据集，并为其补充合成样本。

为了评估 CGMM 相较于其他合成基因组生成模型的表现 [27,36,37,73], ，我们进行了三项实验，旨在生成具有特定属性的基因组（实验1–3）。为此，我们使用了两个数据集：一个包含欧洲人和非洲人占主导的祖源群体，另一个包含健康个体和皮肤黑色素瘤患者。在所有实验中，如表2所示，CGMM 的表现均优于其他模型，能够以更高的转换率将对照 样本转化为可靠的合成病例 样本。

我们重复了每项实验，并纳入了以单核苷酸多态性形式表示的外部知识。结果表明，整合知识进一步提升了 CGMM 的表现，但提升幅度远小于其他所有模型（见表2和图2）。因此，在整合知识后，CGMM 与其他最先进模型的结果相当。据此，我们认为：1）结果表明，与其他模型不同，CGMM 能够学习对照 集与病例集之间的正确关系。因此，整合知识对其影响相对较小。2）CGMM 似乎能够部分捕捉单核苷酸多态性特征。这一结果令人鼓舞，因为经过适当调整后，CGMM 也可用于检测单核苷酸多态性特征。3）即使在小型数据集上，如果没有单核苷酸多态性特征方面的知识，CGMM 也可能取得比其他模型更好的结果。这是因为 CGMM 的突变路径方法能够通过将病例 样本与多个对照 样本配对，将训练集扩展到规模大得多的队列。进而，这增加了可用于训练 CGMM 的突变过程数量，使合成病例 样本的推断更加容易。

此外，图 2 显示，需要在扩大训练集（这有助于提高转换率）与保留样本用于生成合成样本之间取得平衡。因此，我们表明，尽管实验1实现了更高的转换率，但实验2实现的样本数量绝对增量更大（也就是说，尽管实验2使用的训练样本数据集规模更小，但其生成的可靠合成样本比实验1多 31 个）。

本研究存在若干局限性，未来工作应加以解决，以进一步提升 CGMM 的性能和可用性。首先，CGMM 中的自编码器组件旨在优化编码任务，该任务采用自动深度学习方法获得，但未利用任何可能有所帮助的领域知识 [74]。其次，当前方法缺乏可解释性，这可能限制其在临床领域的应用 [75, 76]。第三，由于缺乏可免费获取的基因组学数据，实验是在规模相对较小的数据集上开展的。尽管这些实验涵盖了重要的使用配置，但在统计规模更大的数据集上进行测试，有助于对 CGMM 进行更稳健的评估。

声明

资金支持

本研究未获得公共、商业或非营利部门资助机构提供的专项资助。

利益冲突/竞争性利益

作者　　对本文讨论的任何材料均无财务或专有利益　　e.

数据可用性

所用数据均可公开获取，相关来源已在正文中注明。

作者贡献

Teddy Lazebnik：概念化、数据整理、正式分析、调查、方法学、软件、可视化、监督、撰写初稿以及撰写审阅与编辑。

Liron Simon-Keren：正式分析、可视化和初稿撰写。

致谢
作者谨此感谢 Olga Chervova、Vitaly Voloshin、Ismail Moghul 和 Stephan Beck 对本项目的帮助。

参考文献

[1] C. M. Fraser. 值得庆贺的基因组。《科学》，371(6529):545，2021年。[2] A. J. Gates、D. M. Gysi、M. Kellis 和 A-L. Barabasi。基于人类基因组计划取得的丰硕发现——数字一览。《自然》，590:212–215，2021年。[3] F. S. Alkuraya。人类基因组如何改变罕见疾病研究。《美国人类遗传学杂志》，97(2):199–215，2015年。[4] J. Berg、M. Khoury 和 J. Evans。在临床实践和公共卫生中部署全基因组测序：逐步应对挑战。《遗传医学》，13:499–504，2011年。[5] S. L. Poon、M. N. Huang、Y. Choo、J. R. McPherson、W. Yu、H. L. Heng、A. Gan、S. S. Myint、E. Y. Siew、L. D. Ler、L. G. Ng、W-H. Weng、C-K. Chuang、J. S. P. Yuen、S-T. Pang、P. Tan、B. T. Teh 和 S. G. Rozen。突变特征表明马兜铃酸与膀胱癌发展有关。《基因组医学》，7:38，2015年。[6] L. B. Alexandrov 和 M. R. Stratton。突变特征：隐藏在癌症基因组中的体细胞突变模式。《遗传学与发育学最新观点》，24:52–60，2014年。[7] S. Nik-Zainal、L. B. Alexandrov、D. C. Wedge、P. Van Loo、C. D. Greenman、K. Raine、D. Jones、J. Hinton、J. Marshall、L. A. Stebbings、A. Menzies、S. Martin、K. Leung、L. Chen、C. Leroy、M. Ramakrishna、R. Rance、K. W. Lau、L. J. Mudie、I. Varela、D. J. McBride、G. R. Bignell、S. L. Cooke、A. Shlien、J. Gamble、I. Whitmore、M. Maddison、P. S. Tarpey、H. R. Davies、E. Papaemmanuil、P. J. Stephens、S. McLaren、A. P. Butler、J. W. Teague、G. Jonsson、J. E. Garber、D. Silver、P. Miron、A. Fatima、S. Boyault、A. Langerod、A. Tutt、J. W. M. Martens、S. A. J. R. Aparicio、A. Borg、A. V. Salomon、G. Thomas、A-L. Borresen-Dale、A. L. Richardson、M. S. Neuberger、P. A. Futreal、P. J. Campbell 和 M. R. Stratton。塑造21种乳腺癌基因组的突变过程。《细胞》，149(5):979–993，2012年。[8] E. Birney、T. J. Hudson、E. D. Green、C. Gunter 等。预发表数据共享。《自然》，461:168– 170，2009年。[9] M. Walport 和 P. Brest。共享研究数据以改善公共卫生。《柳叶刀》，377:537–539，2011年。[10] K. Powell。破坏人类基因组研究的承诺落空。《自然》，590:198–201，2021年。[11] O. Chervova、L. Conde、J. A. Guerra-Assuncao 等。英国个人基因组计划：一个开放获取的人类多组学数据资源。《科学数据》，6:257，2019年。[12] J-S. Lee。探索癌症基因组图谱计划的癌症基因组数据。《*BMB*在线报告》，49(11):607–611，2016年。[13] J. Zhang、J. Baran、A. Cros、J. M. Guberman、S. Haider、J. Hsu、Y. Liang、E. Rivkin、J. Wang、B. Whitty、M. Wong-Erasmus、L. Yao 和 A. Kasprzyk。国际癌症基因组联盟数据门户——癌症基因组学数据的一站式平台。《数据库》，2011，2011年。[14] D. Marbach、T. Schaffter、C. Mattiussi 和 D. Floreano。生成逼真的计算机模拟基因网络，用于评估逆向工程方法的性能。《计算生物学杂志》，16(2):229–239，2009年。

[15] T. Mailund、M. H. Schierup、C. N. Pedersen、P. J. M. Mechlenborg、J. N. Madsen 和 L. Schauser。CoaSim：一种基于合并模型模拟遗传数据的灵活环境。*BMC*生物信息学，6:252，2005年。[16] P. F. O'Reilly、L. J. M. Coin 和 C. J. Hoggart。invertFREGENE：用于模拟群体遗传数据中倒位的软件。生物信息学，26(6):838–840，2010年。[17] A. Pinna、N. Soranzo、I. Hoeschele 和 A. de la Fuente。使用 SysGenSIM 模拟系统遗传学数据。生物信息学，27(17):2459–2462，2011年。[18] P-R. Loh、G. Tucker、B. K. Bulik-Sullivan、B. J. Vilhjalmsson、H. K. Finucane、M. Salem、R、D. I. Chasman、P. M. Ridker、B. M. Neale、B. Berger、N. PAtterson 和 A. Price。高效的贝叶斯混合模型分析可提高大型队列中的关联检验效能。自然遗传学，47(3):284–290，2015年。[19] S. Yang、J. Wen、S. T. Eckert、Y. Wang、D. J. Liu、R. Wu、R. Li 和 X. Zhan。使用套索回归和置换辅助调参对全基因组关联研究中的遗传变异进行优先排序。生物信息学，36(12):3811–3817，2020年。[20]J. Mbatchou、L. Barnard、J. Backman、A. Marcketta、J. A. Kosmickim、A. Ziyatdinov、C. Benner、C. O'Dushlaine、M. Barber、B. Boutkov、L. Habegger、M. Ferreira、A. Baras、J. Reid、G. Abecasis、E. Maxwell 和 J. Marchini。用于定量性状和二元性状的高效全基因组回归。自然遗传学，2021年。[21] E. Sawyer、M. Banuelos、R. F. Marcia 和 S. Sindi。用于基因组信号异常检测的神经网络方法。*2020*年亚太信号与信息处理协会年度峰会与会议，第968–971页，2020年。[22] M. Nicolau、R. Tibshirani、A-L. Borresen-Dale 和 S. S. Jeffrey。疾病特异性基因组分析：识别病理生物学特征。生物信息学，23(8):957–965，2007年。[23] S-J. Wang 和 R. Simon。基因组特征在肿瘤学及其他疾病治疗药物开发中的应用。《药物基因组学杂志》，6(8):166–173，2006年。[24] Li. C. 和 M. Li。Gwasimulator：一种快速的全基因组模拟程序。生物信息学，24(1):140–142，2008年。[25] Su. Z.、J. Marchini 和 P. Donnelly。HAPGEN2：多疾病单核苷酸多态性的模拟。生物信息学，27(16):2304–2305，2008年。[26] H. F. Porter 和 P. F. O'Reilly。多变量模拟框架揭示多性状全基因组关联研究方法的性能。《科学报告》，7:38837，2017年。[27] H. V. Meyer 和 E. Birney。Phenotypesimulator：一个用于模拟多性状、多位点基因型到表型关系的综合框架。生物信息学，34(17):2951–2956，2018年。[28] P. C. Phillips。上位性——基因相互作用在遗传系统的结构与演化中的关键作用。《自然评论》，9(11):855–867，2008年。[29] D. B. Fogel 和 J. W. Atmar。使用线性系统模拟进化过程，比较遗传算子与高斯突变。《生物控制论》，63:111–114，1990年。[30] D. I. Podolskiy、A. V. Lobanov、G. V. Kryukov 和 V. N. Gladyshev。癌症基因组分析揭示人类衰老的基本特征及其在癌症发展中的作用。《自然通讯》，7:12157，2016年。[31] R. Gao、A. Davis、T. O. McDonald、E. Sei、X. Shi、Y. Wang、P. C. Tsai、A. Casasent、J. Waters、H. Zhang、F. Meric-Bernstam、F. Michor 和 N. E. Navin。三阴性乳腺癌中突发式拷贝数演化与克隆静止。自然遗传学，48:1119–1130，2016年。[32] I. Y. Zhbannikov、K. G. Arbeev 和 A. I. Yashin。cophesim：用于测试新型关联分析方法的综合表型模拟器。*F1000Research*，6(1294)，2017年。[33] M. Shi、D. M. Umbach、A. S. Wise 和 C. R. Weinberg。模拟常染色体基因型，使其具有真实的连锁不平衡并引入遗传效应。*BMC*生物信息学，19，2018年。

[34] W. Yang 和 Gu. C. C. 一种能够模拟复杂疾病高阶上位性的全基因组模拟器。《遗传流行病学》，37(7):686–694，2013年。[35] D. B. Blumenthal、L. Viola、M. List、J. Baumbach、P. Tieri 和 T. Kacprowski。EpiGEN：一种上位性模拟流程。《生物信息学》，36(19):4957–4959，2020年。[36] Y. Tang 和 X. Liu。G2P：一种用于基因型模拟、表型模拟和统计功效评估的全基因组关联研究模拟工具。《生物信息学》，35(19):3852–3854，2019年。[37] D. A. Reidenbach、A. Lal、L. Slim、O. Mosafi 和 J. Israeli。GEPSi：用于模拟全基因组关联研究表型数据的 Python 库。bioRxiv，2021年。[38] R. A. Weinberg。癌症如何发生。《科学美国人》，275(3):62–70，1996年。[39] H. M. Temin。癌症基因的演化是一个由突变驱动的过程1。《癌症研究》，48(7):1697–1701，1988年。[40] A. F. Rubin 和 P. Green。癌症基因组中的突变模式。《美国国家科学院院刊》，106(51):21766–21770，2009年。[41] A. Makohon-Moore 和 C. A. Iacobuzio-Donahue。从进化视角看胰腺癌的生物学与遗传学。《自然评论：癌症》，16:553–565，2016年。[42] M. J. Fusco、H. J. West 和 C. M. Walko。肿瘤突变负荷与癌症治疗。《JAMA肿瘤学》，7(2):316–316，2021年。[43]C. U. Blank、J. B. Haanen、A. Ribas 和 T. N. Schumacher。癌症免疫图谱。《科学》，352(6286):658–660，2016年。[44] N. Bellomo 和 M. Delitala。从数学动力学和随机博弈论到癌细胞突变、发生、进展及免疫竞争的建模。《生命物理评论》，5(4):183–206，2008年。[45] J. H. Holland。遗传算法。《科学美国人》，267(1):66–73，1992年。[46] L. Davis。将自适应算法应用于上位性领域。《人工智能国际联合会议论文集》，第162–164页，1985年。[47] Z. W. Bo、L. Z. Hua 和 Z. G. Yu。利用遗传算法优化工艺路线。《机器人与计算机集成制造》，22:180–188，2006年。[48] M. Salehi 和 A. Bahreininejad。利用混合遗传算法和智能搜索优化作业车间加工的工艺规划。《智能制造杂志》，22(4):643–652，2011年。[49] L. Bo 和 L. Rein。Luus–Jaakola 优化程序与遗传算法的比较。《工程优化》，37(4):381–396，2005年。[50] A. Ghaheri、S. Shoar、M. Naderan 和 S. S. Hoseini。遗传算法在医学中的应用。《阿曼医学杂志》，30(6):406–416，2005年。[51] J. Zhao 和 M. Xu。利用遗传算法优化阿特金森循环发动机的燃油经济性。《应用能源》，105:335–348，2013年。[52] R. J. Parsons、S. Forrest 和 C. Burks。遗传算法、算子与 DNA 片段组装。《机器学习》，21:11–33，1995年。[53] S. Shah 和 A. Kusiak。利用数据挖掘和遗传算法搜索癌症基因。《生物与医学中的计算机》，37(2):251–261，1995年。[54] S. Katoch、S. S. Chauhan 和 V. Kumar。遗传算法综述：过去、现在与未来。《多媒体工具与应用》，80:8091–8126，2021年。

[55] D.S. Weile 和 E. Michielssen。遗传算法优化在电磁学中的应用：综述。*IEEE* 天线与传播汇刊，45(3):343–353，1997年。[56] T. Bhoskar、O. K. Kulkarni、N. K. Kulkarni、S. L. Patekar、G. M. Kakandikar 和 V. M. Nandedkar。遗传算法及其在机械工程中的应用：综述。今日材料：会议录，2(4):2624–2630，2015年。[57]D. B. Hibbert。遗传算法在化学中的应用。化学计量学与智能实验室系统，19(3):277–293，1993年。[58] S.K. Pal、S. Bandyopadhyay 和 S.S. Ray。进化计算在生物信息学中的应用：综述。*IEEE* 系统、人类与控制论汇刊 *C*辑（应用与综述），36(5):601–615，2006年。[59] B. Chowdhury 和 G. Garai。从遗传算法视角对多序列比对进行综述。基因组学，109(5):419–431，2017年。[60] B. D. Ondov、P. Treangen、T. J. Melsted、A. B. Mallonee、N. H. Bergman、S. Koren 和 A. M. Phillippy。Mash：利用最小哈希快速估计基因组和宏基因组距离。基因组生物学，17:132，2016年。[61] A. Shukla、H. M. Pandey 和 D. Mehrotra。遗传算法中选择技术的比较综述。载于*2015*年计算分析与知识管理未来趋势国际会议（*ABLAZE*），第515–519页，2015年。[62] D. Dvinskikh、A. Ogaltsov、A. Gasnikov、P. Dvurechensky、A. Tyurin 和 V. Spokoiny。适用于凸与非凸随机优化的自适应梯度下降。*arXiv*，2020年。[63] H. Jin、F. Chollet、Q. Song 和 X. Hu。AutoKeras：用于深度学习的自动化机器学习库。机器学习研究杂志，24(6):1–6，2023年。[64] J. Wu、X-Y. Chen、H. Zhang、L-D. Xiong、H. Lei 和 S-H. Deng。基于贝叶斯优化的机器学习模型超参数优化。电子科技大学学报，17(1):26–40，2019年。[65] F. Almeida 和 G. Xexeo。词嵌入：综述。*arXiv*，2019年。[66] A. F. Ganai 和 F. Khursheed。使用循环神经网络和长短期记忆单元预测下一个词：统计语言建模。载于*2019*年第五届图像信息处理国际会议（*ICIIP*），第469–474页，2019年。[67] O. Chervova、L. Conde、J. A. Guerra-Assuncao、I. Moghul、A. P. Webster、A. Berner、E. L. Cadieux、Y. Tian、V. Voloshin、T. F. Jesus、R. Hamoudi、J. Herrero 和 S. Beck。英国个人基因组计划：一个开放获取的人类多组学数据资源。《科学数据》，6:257，2019年。[68] J. D. Churchill、S. E. Schmedes、J. L. King 和 B. Budowle。评估 Illumina® beta 版 ForenSeq™ DNA Signature Prep Kit 用于遗传分型。国际法医科学：遗传学，20:20–29，2016年。[69] J. Alexandrov、L. B. Kim、N. J. Haradhvala、M. N. Huang、A. W. T. Ng、Y. Wu、A. Boot、K. R. Covington、D. A. Gordenin、E. N. Bergstrom、S. M. A. Islam、N. Lopez-Bigas、L. J. Klimczak、J. R. McPherson、S. Morganella、R. Sabarinathan、D. A. Wheeler、V. Mustonen、G. Getz、S. G. Rozen、M. R. Stratton 和 PCAWG联盟。人类癌症中的突变特征谱。《自然》，578:94–101，2020年。[70] D. Defays。一种高效的全连接法算法。计算机杂志，20(4):364–366，1977年。[71] M. L. Menendez、J. A. Pardo、L. Pardo 和 M. C. Pardo。詹森-香农散度。富兰克林研究所学报，334(2):307–318，1997年。[72] J. M. Joyce。库尔贝克-莱布勒散度，第720–722页。施普林格·柏林·海德堡，2011年。[73] J. Zhou、P. Zhong 和 T. Zhang。一种基于复杂网络特征的无比对DNA序列相似性分析新方法。进化生物信息学，12:229–235，2016年。[74] T. Lazebnik 和 S. L. Keren。知识集成自编码器模型。*arXiv*，2023年。

[75] Y. A. Veturi、W. Woof、T. Lazebnik、I. Moghul、P. Woodward-Court、S. K. Wagner、T. A. Cabral de Guimaraes、M. D. Varela、B. Liefers、P. J. Patel、S. Beck、A. R. Webster、O. Mahroo、P. A. Keane、M. Michaelides、K. Balaskas 和 N. Pontikos。Syntheye：探究合成数据对人工智能辅助遗传性视网膜疾病基因诊断的影响。*Ophthalmology Science*，第 100258 页，2022 年。

[76] T. Lazebnik 和 S. Bunimovich-Mendrazitsky。使用 SAT-PP算法进行决策树后剪枝且不损失准确率：基于临床数据的实证评估。*Data & Knowledge Engineering*，145:102173，2023 年。

附录

所提出模型的超参数说明及取值已用于本文所述实验。这些取值是在开发 CGMM 的试错过程中，针对本文实验 1 所用样本的一个子集选定的。

参数定义	**取值**
配对方法 [1]	子集方法，使用每种情况的 10 对样本
RBGA 中的代数 [1]	100
遗传算法中的初始变异率 [1]	$2.5 \cdot 10^{-4}$
遗传算法中的变异率递减因子 [1]	$1 \cdot 10^{-5}$
遗传算法中的特许权使用费率 [1]	0.05
停止条件 [1] 的阈值 ζ	$1.5 \cdot 10^{3}$
加速有界变异算子 [1] 的界定因子 β	$4.5 \cdot 10^{3}$
自编码器的编码层维度 [1]	8192

表 3：模型超参数及其取值说明。