

群体遗传学模拟软件应用现状

高峰^{1,2}, 李海鹏¹

1. 中国科学院计算生物学重点实验室, 中国科学院-德国马普学会计算生物学伙伴研究所, 上海 200031;
2. 中国科学院大学, 北京 100049

摘要: 随着下一代测序技术的不断进步与测序价格的不断下降, 越来越多物种的全基因组信息被公开。作为研究群体遗传变异模式工具之一的模拟软件必然将发挥越来越重要的作用。依据时间推演方向的不同, 模拟软件可以分为依时间向前和向后推演, 二者各有所长, 功能上互相补充, 分别适合于不同的模拟需求。这些软件在研究进化动力的影响、估计进化动力参数与验证不同进化假设以及新方法有效性等方面起着重要作用。本文简要介绍了群体遗传学相关理论知识, 详细比较了近 10 年来发表的 32 款模拟软件, 并对模拟软件的未来发展方向给出了建议。

关键词: 群体遗传学; 计算机模拟; 依时间向前; 依时间向后

Application of computer simulators in population genetics

Feng Gao^{1,2}, Haipeng Li¹

1. CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The genomes of more and more organisms have been sequenced due to the advances in next-generation sequencing technologies. As a powerful tool, computer simulators play a critical role in studying the genome-wide DNA polymorphism pattern. Simulations can be performed both forwards-in-time and backwards-in-time, which complement each other and are suitable for meeting different needs, such as studying the effect of evolutionary dynamics, the estimation of parameters, and the validation of evolutionary hypotheses as well as new methods. In this review, we briefly introduced population genetics related theoretical framework and provided a detailed comparison of 32 simulators published over the last ten years. The future development of new simulators was also discussed.

Keywords: population genetics; computer simulation; forwards-in-time; backwards-in-time

收稿日期: 2016-03-22; 修回日期: 2016-04-25

基金项目: 中国科学院先导 B 项目(编号: XDB13040800)和国家自然科学基金项目(编号: 91531306)资助[Supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13040800) and the National Natural Science Foundation of China (No. 91531306)]

作者简介: 高峰, 博士, 专业方向: 生物信息学。E-mail: gaofeng@picb.ac.cn

通讯作者: 李海鹏, 博士, 研究员, 研究方向: 理论群体遗传学。E-mail: lihaipeng@picb.ac.cn

DOI: 10.16288/j.yczs.16-100

网络出版时间: 2016/7/25 12:47:36

URI: <http://www.cnki.net/kcms/detail/11.1913.R.20160725.1247.001.html>

群体遗传学(Population genetics)诞生于 20 世纪初期,是一门研究物种在群体水平受各种进化动力,如突变、遗传漂变、遗传重组、自然选择和群体历史(包括迁移和瓶颈效应等)造成的等位基因频率与基因型频率随世代变化规律的遗传学分支学科^[1]。为了研究物种极其复杂的进化过程,理论研究者借助于数学和统计学方法提出了多种进化模型。这些模型通常有着特定的前提假设,通过简化复杂的进化过程,使人们得以研究群体遗传变异随时间的变化规律。

与传统理论方法的繁杂数学推导不同,在设定计算机模拟软件(Computer simulator)的参数之后,人们就可以模拟物种在各种进化动力和不同群体遗传结构影响下的长期进化过程,从而获得大量的模拟数据。所以计算机模拟对理论建模是一个良好的补充,为不同领域的学者研究物种复杂的进化过程提供了有力工具。越来越多模拟软件的发表也说明了计算机模拟在群体遗传学中扮演的重要性^[2-4]。总的来说,群体遗传学模拟软件主要有以下作用:

(1) 预测与统计推断。预测指的是观察多种进化动力对群体遗传变异模式造成的影响。通过初始化物种群体祖先信息,设定群体经历的进化动力参数(如迁移率和交配机制等),经过若干世代的模拟就可以得到群体当代的遗传变异模式^[5-7]。在已知现生群体观测数据的基础上,人们可以通过比较计算机模拟数据和观测数据的差异来推断群体所经历的复杂进化过程^[8,9]。特别是渐进贝叶斯计算(Approximate Bayesian computation, ABC)^[10]方法,通过在观测数据相关统计量已知先验分布中抽样并产生相应的模拟数据,基于这些模拟数据就可以获得所研究参数的后验分布。

(2) 验证新方法的有效性。群体遗传学各个研究方向不断涌现出新的方法与模型(如检验正选择^[11]和估计群体遗传重组率^[12,13]等),这些新方法的提出促进了相关研究的不断深入,对学科发展具有重要意义。另一方面,随着高通量测序技术的不断发展,人们可以很方便地获得所研究物种的基因组信息,甚至是群体水平的信息。然而,这些基因组水平数据是不能用来检验一个新方法的有效性的。这是因为这些物种的基因组数据相当复杂,不但数据的产

生过程复杂,而且最重要的是这些物种经历过的历史是未知的,而且不同染色体区域受到各种进化作用的影响往往是不一样的,比如说突变速率、遗传重组率、或者受到选择与否。但是模拟数据的产生过程是具备良好的已知前提的,因此模拟软件产生的(基因组水平)数据可以用来验证这些新方法的运算效率与准确性^[11,12,14]。

(3) 群体遗传学教学等其他应用。与传统的文本教学方式相比,具有图形操作界面的模拟软件更加生动、活泼。通过与软件交互,实际动手操作软件,学生们更能快速深刻理解相关的遗传学知识,如孟德尔遗传定律、PCR 实验流程等等^[15-17]。而且模拟软件也可用于辅助实验,在实验设计初期,模拟数据可以评估达到预期统计显著性所需要的资源^[18],实验后期可用于分析统计不显著结果等^[19]。此外,对全基因组数据的模拟也可用于全基因组关联分析(Genome wide association study, GWAS)^[20]。最后,模拟软件在保护遗传学^[6,21]与流行病学^[22,23]研究中也有着广泛的应用。

近几年发表的几篇综述文献评估了当时流行的模拟软件^[24-27]。在此基础上,本文首先简要阐述了群体遗传学相关理论知识、两类模拟软件的优缺点与适用范围,然后详细介绍了各个模拟软件的可模拟情景,最后对模拟软件的未来发展趋势给出了建议。

1 理论基础

1.1 中性 Wright-Fisher 模型

1968 年, Kimura^[28]提出了著名的中性理论(Neutral theory),认为大多数突变都是中性的,不会影响个体的适应度,因此个体也就不会受自然选择影响。在此框架下,遗传漂变是主要的进化动力。遗传漂变指的是群体产生下一代的过程中由于抽样误差导致的等位基因频率的随机波动。当群体较小时,遗传漂变作用将会变强。中性突变产生新的等位基因,受遗传漂变影响,这些等位基因在群体中逐步消失或者固定下来,因此中性进化理论也称为突变-漂变假设(Mutation-drift hypothesis)。该理论成功地解释了物种内或者物种间观察到的遗传变异现象。现在,中性进化理论已经成为检验选择是否发生的原假设,也就是说,如果预期目标是检验选择,那么就需要

拒绝中性假设。

在中性理论的基础上, 发展出来的 Wright-Fisher 模型^[29]是模拟遗传漂变效应最基础的遗传模型。它具有多个前提假设条件: 群体大小恒定不变; 每代之间是离散的; 个体之间随机交配; 不存在自然选择。在产生下一代的过程中, 每个后代个体随机选取一个父代个体, 因此一个父代个体可以拥有很多后代个体。假设群体大小为 $2N$, 那么每个父代个体拥有的后代个体数服从 $n=2N, p=1/2N$ 的二项分布, 当 $1/2N$ 远小于 1 时, 近似服从参数为 $\lambda=1$ 的泊松分布 (Poisson distribution), 也就是每个父代个体拥有的平均后代个体数为 1。

1.2 溯祖理论

在中性 Wright-Fisher 模型的基础上, Kingman^[30]于 20 世纪 80 年代提出了溯祖理论 (Coalescent theory)。顾名思义, 溯祖就是多个样本 (或者染色体) 回溯到最近共同祖先 (Most recent common ancestor, MRCA) 的过程。而回溯过程可以看做为逐步构建系谱树 (Genealogy tree) 的过程。在依时间向后推断时, 两个样本发生一次溯祖事件的结果便是它们找到了这两个样本的最近共同祖先。经过多次溯祖事件, 系谱树逐步构建完成, 树的根叫做所有样本的最近共同祖先。图 1 展示了 4 个个体的溯祖过程, 经过 3 次溯祖事件, 4 个个体找到了它们的最近共同祖先。

系谱树的枝长表示溯祖时间, 溯祖时间与系谱树的拓扑结构是独立不相关的。在图 1 中, 树的枝

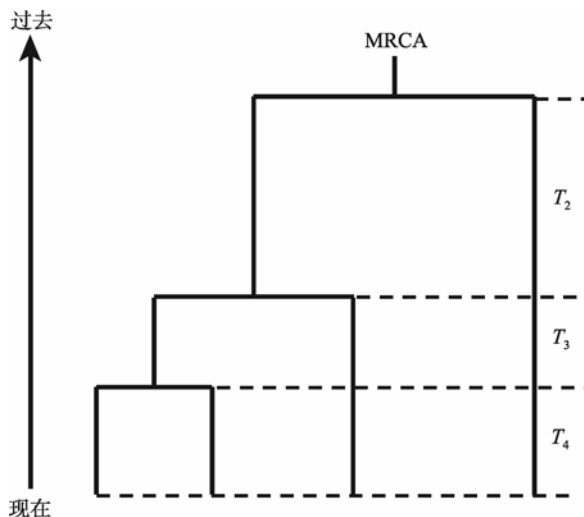


图 1 4 个个体溯祖过程

Fig. 1 The coalescent process of four individuals

长 T_k 表示在当前有 k 个枝的情况下发生下一次溯祖事件的时间, T_k 近似服从指数分布, 其期望

$$E(T_k) = \frac{4N_e}{k(k-1)}, \text{ 其中, } N_e \text{ 表示有效群体大小 (Effective}$$

population size)。系谱树的总枝长为 $L = \sum_{k=2}^n kT_k$, 其

期望为 $E(L) = 4N_e \sum_{i=1}^{n-1} \frac{1}{i}$, 其中, n 表示样本大小。需要

注意的是, 当 N_e 很大的时候, 同一时间发生两次以上的溯祖事件的概率太小了 $\left(\leq \left(\frac{1}{2N_e} \right)^2 \right)$, 所以回溯

的过程中一般不需要考虑这种情况。

无限位点突变模型 (Infinite-site mutation model)^[31]是中性突变的标准模型, 指的是每次突变都会发生在新的位点上。中性突变不会影响树的拓扑结构, 与溯祖过程也是独立的。整个系谱树的总突变数 M 服从 $\lambda = \theta L$ 的泊松分布, 其中, θ 为群体突变率。总

突变数 M 的期望为 $E(M) = \theta \sum_{i=1}^{n-1} \frac{1}{i}$ ^[32]。

另外许多学者对溯祖理论在遗传重组、自然选择和其他复杂群体历史模型方面进行了理论扩展^[32-35]。比如遗传重组事件会将模拟序列打断成几个较短的子片段, 每个子片段拥有各自的系谱树。给定群体遗传重组率 ρ , 那么遗传重组发生次数 R 的期望为

$$E(R) = \rho \sum_{i=1}^{n-1} \frac{1}{i}$$
^[32]。而对于自然选择和群体历史的处理

比较复杂, 在此就不加以阐述了。

溯祖理论以数学的角度分析样本观测数据, 可以快速有效地生成不同进化情景下的模拟数据 (见 2.2 部分), 使人们直观理解不同进化动力对群体遗传变异模式造成的影响。目前溯祖理论已经被广泛应用到群体遗传学多个领域, 如检验选择^[11]、真实群体的群体参数估计^[34]以及疾病基因图谱^[36]等。

2 模拟软件

根据依时间推算的方向不同, 群体遗传学模拟软件可以分为依时间向前 (Forwards-in-time) 模拟软件和依时间向后 (Backwards-in-time) 模拟软件。这两类模拟软件分别基于不同的模拟策略, 二者在可模

拟情景范围、情景复杂程度和运行时间等方面有区别,因此它们分别适合于不同的模拟需求。下面将详细介绍两类模拟软件。

2.1 依时间向前模拟

依时间向前模拟从初始祖先群体出发(通常包括数千个个体),模拟祖先群体中的所有个体在预先设定的进化模型下遵循的出生、选择、交配、生产、迁移和死亡的整个生命周期,以每个世代为单位来观察特定世代区间内的群体遗传结构变化规律。一方面,考察群体所有个体在特定进化情景下的变化使得该策略可以追踪整个群体祖先信息、模拟群体极其复杂的进化过程和复杂的疾病模型、观察任意世代群体相关信息等;另一方面,该模拟策略非常耗时,而且比较耗内存,通常需要模拟数千世代以使基因频率达到平衡。

按软件发表时间先后顺序,表 1 列举了多种依时间向前模拟软件的可模拟情景信息。EASYPOP^[37]是最先发表的模拟软件,实现了距离隔离模型(Isolation-by-distance model)^[38],也就是迁移率随着距离

的增加而指数递减,但其功能有限,仅限于模拟没有自然选择的情景。rmetasim^[39]是用 R 语言编写的,充分利用了 R 语言强大的画图功能。simuPOP^[40]可以模拟各种染色体、群体内部非随机交配等多种交配机制(如植物的不同自交率)、复杂的自然选择模型和用户自定义的疾病相关等位基因频率等。simuPOP 非常灵活,支持运行用户自己编码的模型^[41],其扩展版本^[42]可以模拟复杂疾病对群体多个世代造成的影响。Fregene^[43]可以模拟多个遗传座位受到正负选择或者平衡选择的情景,提供适用于流行病学研究中的确定基因抽样功能等。通过采用在模拟时重新设定模拟参数的策略(如降低有效群体大小数值等),Fregene 具有较快的运行速度,然而该策略在一定程度上降低了模拟数据的精确度^[44]。随后几年发表的 Nemo^[45]、quantiNemo^[46]和 ForSim^[47]也将环境因素纳入模拟范围。前两者支持模拟群体发生随机的灭绝和饥饿事件,支持对不同的群体密度设定不同的迁移率参数,后者支持模拟基于表型(Phenotype)的选择性交配机制。Nemo 可以模拟自然选择与环境在群体水平的交互作用,也可以模拟有害突变。quantiNemo

表 1 依时间向前模拟软件的可模拟情景信息

Table 1 Simulation scenarios of forwards-in-time simulator

软件名称	遗传重组	基因转换	选择	群体大小变化	迁移	基因组数据	疾病模型	编程语言
EASYPOP	是	否	否	否	是	否	否	C
Rmetasim	否	否	否	是	是	否	否	R
simuPOP	是	否	是	是	是	否	是	Python
Nemo	是	否	是	是	是	否	是	C++
Mendel Accountant	变	否	是	是	是	是	否	C++
Fregene	是	是	是	是	是	是	是	C++
quantiNemo	是	否	是	是	是	否	是	C++
ForSim	否	否	是	是	是	否	是	C++
genomeSIMLA	是	否	否	是	否	是	是	C++
SFS_CODE	是	是	是	是	是	否	否	C
GENOMEPOP	是	否	是	否	是	是	否	C++
ForwSim	是	否	是	否	否	是	否	C++
Pedagog	否	否	是	是	是	否	否	VB
AnA-FltS	是	否	是	是	是	是	否	C++
Vortex	否	否	是	是	是	是	是	C++
Forqs	是	否	是	是	是	是	否	C++

注:各种情景中,“是”表示可以模拟该情景;“否”表示不可以模拟该情景。遗传重组情景中“变”代表模拟变化的遗传重组率。

研究选择、突变、遗传重组和遗传漂变对数量性状(Quantitative traits)造成的影响。genomeSIMLA^[48]可以模拟人群的真实连锁不平衡(相邻位点间的非随机关联)模式(Linkage disequilibrium, LD),这对于进行复杂疾病的 GWAS 分析是非常最要的。SFS_CODE^[49]支持模拟包括插入与删除(Insertions and deletions)在内的多种突变类型。GENOMEPOP^[50]也采用了重新设定模拟参数策略,因此它也具有较快的模拟速度,此外该软件可以模拟复杂的核苷酸密码子模型,正负选择和任意群体迁移模型。Yuan^[27]评估了 Nemo、quantiNemo 和 GENOMEPOP 这 3 个软件,从理论分析和实际模拟测试角度证明了 GENOMEPOP 的运行速度较快。ForwSim^[51]采取区别对待中性突变与非中性突变的策略来加快模拟速度。由于中性突变不影响个体存活率,ForwSim 采取延迟若干世代的策略来模拟中性突变。由于大部分突变都是中性突变^[52],所以这种模拟策略可以显著地提高模拟效率。通过将一个染色体表示为多个单体型块(Haplotype block)的组合,forqs^[3]以单体型块为基本单位来构造模拟数据,该策略大大降低了运算时间与内存使用率(在没有自然选择情景下,当群体大小为 10000 时,每模拟 100 世代只需要 3 秒)。与溯祖理论构建系谱树的思路相似,AnA-FiTS^[53]提出以后向事件图(Backward event graph, BEG)存储历史事件信息的策略,该策略使得 AnA-FiTS 可以模拟基因组水平数据,而且在运算速度和内存使用率方面也有极大改善。

另外, simuPOP 和 Nemo 都支持模拟拷贝数变异(Copy number variation, CNV)数据和微卫星(Microsatellite)数据。simuPOP 和 Mendel Accountant^[54]都支持利用集群进行并行模拟。pedagog^[55]和 Vortex^[56]提供了用户友好的图形操作界面,pedagog 也可以模拟基因分型错误(Genotyping error)和缺失数据(Missing data)。rmetasim 和 Vortex 都支持模拟随机发生的历史事件。历史事件随机发生指的是预先设定的历史事件在特定的时间点是不一定发生的,该属性在模拟群体经历脆弱的环境、爆发疾病或者气候不稳定时非常有用。

需要注意的是,依时间向前模拟策略在开始模拟前需要指定初始世代遗传变异的相关初始状态,

也就是每个个体的基因型信息。但是群体祖先世代的实际数据通常是无法得到的。一个解决办法就是采取随机策略,假设祖先世代等位基因频率服从正态分布或者均匀分布,如 pedagog。另外一个解决办法是与依时间向后模拟策略相结合,如 rmetasim 可以读取 SIMCOAL2^[57]和 fastsimcoal^[58]的输出来设定随机种子。

2.2 依时间向后模拟

依时间向后模拟也叫溯祖模拟。溯祖模拟效率高有两个原因:第一个原因是因为溯祖理论只追踪样本的信息,而不考虑群体中的所有个体。这就像拍电影,导演只考虑镜头之内的情况,镜头之外的情况则视为不存在。另外一个原因是在模拟过程中仅考虑溯祖事件,而两个相邻溯祖事件发生的时间间隔往往可能跨越了多个世代。

由于溯祖模拟没有考虑群体中所有个体的生命周期过程,因此它只能模拟比较简单的群体历史情景和简单的自然选择过程(单基因座位的双等位基因受正选择的情景)。在模拟正选择情景前,首先需要获得受选择等位基因频率随时间的变化规律,也称为轨迹(Trajectory)。通过给定受选择基因的出现时间与当前世代的频率信息,可以通过确定性模型来推断轨迹;另外也可以通过不同的选择模型产生相应的随机轨迹。

需要注意的是,经典的溯祖模拟在模拟较短的序列(<5 Mb)时效率非常高,然而在模拟基因组水平数据时效率相对较低,甚至不能模拟大样本(>500)基因组数据^[59]。这是因为随着模拟序列的增长和时间的回溯,会发生较多的溯祖事件和遗传重组事件,使得溯祖过程构成复杂的 ARG(Ancestral recombination graph),而不是简单的二叉系谱树结构,特别是遗传重组率较高的时候,因此模拟整个 ARG 会相当耗时。Wiuf 和 Hein^[60]首先提出了在有遗传重组事件时以序列化方式(Sequential)逼近真实的溯祖过程。在此基础上,McVean 和 Cardin^[61]提出了 SMC(Sequentially Markov coalescent)算法,该算法从模拟序列的左端系谱树开始逐步向右端移动,在移动的过程中纳入遗传重组事件来逐步更新系谱树。每个遗传重组事件发生时,当前系谱树被打断的分支可

以与系谱树的其他任意分枝发生溯祖事件, 结果产生一个新的系谱树。后续研究表明 SMC 算法与经典 ARG 模型产生的模拟数据在连锁不平衡和多态性模式方面是高度一致的^[62]。在 SMC 算法的基础上, 其他学者提出了其改进算法 SMC'^[63]和 MaCS(Markovian coalescent simulator)^[64]。SMC' 算法允许被打断的分枝重新跟被打断前的分枝发生溯祖事件, MaCS 算法被认为是推广的 (Generalized) SMC。在保留 SMC 算法模拟速度快和内存使用率低的优点上, SMC' 和 MaCS 证明比 SMC 更逼近真实的溯祖过程^[63,64]。

按软件发表时间先后顺序, 表 2 列举了多种依时间向后模拟软件的可模拟情景信息。最经典的也是目前使用最多的是 Hudson^[65]编写的 ms。在中性 Wright-Fisher 模型的假设下, ms 假定中性突变服从无限位点突变模型。此外, ms 可以模拟遗传重组, 基因转换, 亚群体间对称迁移和瓶颈效应、群体扩张等多种群体历史事件。ms 的标准输出包括多态性数据与每个分离位点的相对位置, 也可以输出系谱树。msHOT^[66]和 Mlcoalsim^[67]分别对 ms 进行了扩展。msHOT 可以模拟任意位置和任意强度的遗传重组

热点 (Recombination hotspots, 一段序列的遗传重组率高于附近区间的遗传重组率) 和基因转换。在多个进化模型假设下, Mlcoalsim 可以生成一个或者多个遗传座位 (独立或者连锁) 数据, 并对这些数据进行包括 Tajima's D 在内的多个统计检验。GENOME^[68]和 SIMCOAL2 以离散的每个世代为单位来逐步构造系谱树, 另外, 前者允许多个个体溯祖到同一祖先, 该方法可以较快的模拟基因组水平数据, 后者允许多个溯祖事件发生在同一世代上。SIMCOAL2 可以模拟更加复杂的历史情景, 但是世代隔世代 (Generation by generation) 的策略比较耗时。由于物种的进化是一个极其复杂的过程, SPLATCHE^[69]首次考虑环境因素对遗传变异模式造成的影响, 这需要用户输入物种群体相关的地理信息系统 (Geographic information system, GIS) 信息来定义环境抵抗力。CoaSim^[70]和 MODELER4SIMCOAL2(m4s2)^[71]提供了简单易用的图形用户操作界面。Serial SimCoal^[72]可以模拟多个群体在过去多个时间点的遗传信息。这些模拟数据类似于在物种化石或者标本中提取的真实数据。

表 2 依时间向后模拟软件的可模拟情景信息

Table 2 Simulation scenarios of backwards-in-time simulator

软件名称	遗传重组	基因转换	选择	群体大小变化	迁移	基因组数据	编程语言
ms	是	是	否	是	是	否	C
SPLATCHE	否	否	否	是	是	否	C++
SeiSim	变	否	是	否	否	否	C++
SIMCOAL2	变	否	否	是	是	否	C++
CoaSim	变	是	否	是	是	否	Python
Serial SimCoal	否	否	否	是	是	否	C++
msHOT	变	是	否	是	是	否	C
Mlcoalsim	变	否	是	是	是	否	C
m4s2	变	否	否	是	是	否	Java
GENOME	变	否	否	否	是	是	C++
MaCS	变	是	否	是	是	是	C++
mbs	变	否	是	是	是	否	C
msms	变	否	是	是	是	否	Java
fastsimcoal	变	否	否	是	是	是	C++
fastsimcoal2	变	否	否	是	是	是	C++
Cosi2	变	是	是	是	是	是	C++

注: 各种情景中, “是”表示可以模拟该情景; “否”表示不可以模拟该情景。遗传重组情景中“变”代表模拟变化的遗传重组率。

MaCS 和 fastsimcoal 都是基于改进的 SMC 算法的, 在有遗传重组热点的情景下, 这两个模拟软件比其他模拟软件能更快速灵活的模拟多种历史情景。fastsimcoal2^[73]是 fastsimcoal 的更新版本, 它加入了通过突变频谱(Mutation frequency spectrum, MFS)来估计群体历史参数的功能。fastsimcoal2 也实现了渐进贝叶斯计算方法, 因此可以从预先已知的先验分布中抽样来估计未知群体历史参数的后验分布。

在模拟自然选择方面, SelSim^[74]可以模拟在有遗传重组情景下的正选择和平衡选择。mbs^[75]也是对 ms 的扩展, 它需要用户输入受选择等位基因频率变化轨迹和群体大小变化信息。除了包含 ms 的所有功能外, msms^[76]可以模拟拥有群体结构的选择过程。通过建立部分 ARG, 使用段列表(Segment list)等策略对溯祖过程进行优化, Cosi2^[4]可以在基因组水平模拟正选择。另外, Cosi2 还支持群体结构, 群体大小变化, 迁移, 遗传重组热点和基因转换等情景。通过实际模拟比较 Cosi2 比 msms 和 mbs 模拟效率更高^[4]。

通过对 ms、msHOT、MaCS 和 GENOME 四个软件的时间复杂度进行理论分析, 并加上实际模式测试, 发现 GENOME 的运行速度是最快的^[27]。Yang 等^[59]在运行时间和遗传重组热点情景下模拟数据的准确性两个方面评估了 ms、msHOT、MaCS、SIMCOAL2 和 fastsimcoal 五个模拟软件, 发现当遗传重组热点次数增加时, msHOT 的运行时间会变长, 且此时的模拟数据精度不高。而 fastsimcoal 是运行速度最快, 遗传重组热点次数较多时模拟数据最可靠的软件。

2.3 软件使用方式与编程语言

当用户需要使用模拟软件生成模拟数据时, 首先要区分选择依时间向前模拟软件还是选择依时间向后模拟软件。如果用户关注的是群体本身复杂的进化过程, 而不太关注群体进化的结果, 那么就可以选择依时间向前模拟软件, 反之则选择依时间向后模拟软件。然后根据表 1 或者表 2 中列举的软件可模拟情景选择合适的模拟软件。

本文共列举了 32 个模拟软件, 其中有 29 个模拟软件的运行方式是命令行运行, 用户可以在参数设置文本文件或者命令行参数中设置软件运行参数。另外一种运行方式是图形用户界面运行, 用户可以

通过文本框、下拉列表框或者点击按钮等方式设置软件运行参数。命令行运行方式的优点是利于批处理运行以便生成大量模拟数据, 方便与其他软件集成, 并且可以跨平台运行。通过文本文件来设置参数的方式也非常灵活, 非常适合用户需要多次运行软件但是参数变动不大的情形。图形用户操作界面运行方式简单, 用户通过简单的鼠标操作就可以完成参数设置, 非常适合对软件不熟悉的初学者。其中一些软件还提供互动式的图形展示结果的功能(如 Vortex 和 SPLATCHE 等)。

本文列举的 32 个软件中, 共有 27 个是用 C/C++ 语言编写的, 充分利用了 C/C++ 语言灵活高效和较高移植性的特点。其他的软件, 如 msms 和 m4s2, 使用面向对象编程语言 Java 编写的; rmetasim 使用 R 语言编写, 利用了 R 语言强大的图形绘制函数。simuPOP 和 CoaSim 是使用脚本语言 Python 编写。

3 结语与展望

下一代测序技术的不断发展使得越来越多物种的全基因组数据信息被公开, 这就需要更多的新方法和新假设来分析这些数据, 因此基因组水平的模拟数据也会变得非常重要。所以作为研究群体遗传变异模式有力工具之一的模拟软件将在群体遗传学的研究中发挥着越来越重要的作用。现在已经公开发表多达几十个模拟软件供其他研究人员使用。这些程序可以用于研究各种进化动力对遗传变异模式造成的影响, 统计推断遗传模型参数和评估新方法的有效性等。模拟策略归结为依时间向前模拟和依时间向后模拟两种方式。理论上说, 依时间向前模拟策略可以灵活地模拟包括复杂疾病模型在内的任意群体历史情景, 但由于需要追踪群体完整的祖先信息, 故其模拟效率低; 相反的, 只关注与样本有关的祖先信息使得依时间向后模拟策略的模拟效率非常高, 尤其是基因组水平数据的模拟, 但支持模拟的群体历史情景和选择情景都相对简单。因此, 两种模拟策略功能上互相补充, 分别适用于不同的模拟需求。

关于模拟软件的发展方向, 有学者提出将两种模拟策略结合在一起的方法^[51], 但是该方法丢失了依时间向前模拟的特点。结合本课题组长期的研究

经验,对模拟软件的未来发展方向提出 3 点建议:

(1) 模拟的灵活性与高效率。灵活性指的是模拟软件可以在基因组水平模拟多个时间点^[72]、多物种群体^[77]或者包括生态环境在内的多种复杂的进化模型^[78]数据。同时要求模拟运行速度快,内存使用率低,也就是要提高模拟效率,满足生成大量模拟数据的需求。然而二者是互相矛盾的,当模拟模型越复杂时,模拟效率会降低,因此如何在保证模拟高效率的同时保证灵活性,是今后模拟软件的一个发展方向。虽然 SMC 算法及其改进算法提高了溯祖模拟的模拟效率,但是随着高通量测序技术的高度发展,分子生物学领域已经进入大数据时代^[79],这就要求模拟软件能够模拟出与测序技术相关的基因分型错误、偏差或者缺失数据等在内的基因组水平数据。因此,这就要求理论群体遗传学者能够提出更有效的算法来生成相应的模拟数据。

(2) 与其他数据分析软件的集成。除了本文论述的模拟软件外,在群体遗传学领域也存在着许多流行的数据分析软件,如 Arlequin^[80]和 MEGA^[81]等。它们实现了群体遗传学中基本的分析与统计检验方法,可以方便快速地处理遗传学数据。因此,如果这些数据分析软件可以直接读取模拟软件的输出数据,那么将会加快整个分析流程,也间接地扩展了模拟软件的功能。例如,ms 输出的系谱树文件可以作为支持多种进化模型的 seq-gen^[82]的输入,这样可以生成核苷酸序列多态性数据。Carvajal-Rodríguez^[25]提出以软件工程的角度来设计开发模拟软件,要求软件开发者提供详细的需求分析文档,这对于模拟软件的设计开发、操作和维护都有着重要意义。因此,模拟软件与其它数据分析软件的集成将会极大地简化数据的分析过程。

(3) 良好的用户体验。任何软件的开发目的之一是为了让用户方便地解决相关问题。因此,软件的易用性是至关重要的。开发者发布软件时也需要发布详细的用户使用手册与示例程序,以方便用户快速掌握软件使用方法。命令行运行方式是目前绝大多数模拟软件的运行方式,它的优点是可以批处理生成大量数据,而它的缺点是需要用户详细阅读软件操作手册,熟悉相关的参数设置方法。有些软件(如 CoaSim 等)还需要额外的用户编码来模拟更复杂

的模型。这些问题对于大部分用户来说是非常棘手的。因此,图形用户操作界面运行方式非常适合没有相关使用经验的用户。集群、分布式计算系统和云计算等计算机应用技术也为模拟软件提供了发展平台。云计算作为一个新兴概念为处理大数据提供了一套行之有效的解决办法。其服务模式之一软件即服务(Software as a service, SaaS)指的是将应用作为服务提供给用户,用户可以在多种客户端设备,如电脑和手机上方便的使用应用。如果将云计算应用到群体遗传学模拟软件上,人们可以通过移动设备随时操作软件,省去了软件的安装步骤,并借助云设备提供商的硬件支持,极大的提高了运行效率。可以遇见,云计算将会广泛地应用到科学研究中。

参考文献(References):

- [1] Hartl DL, Clark AG. Principles of population genetics. 4th ed. Sunderland, Mass: Sinauer Associates, 2007. [DOI]
- [2] Messer PW. SLiM: simulating evolution with selection and linkage. *Genetics*, 2013, 194(4): 1037–1039. [DOI]
- [3] Kessner D, Novembre J. Forqs: forward-in-time simulation of recombination, quantitative traits and selection. *Bioinformatics*, 2014, 30(4): 576–577. [DOI]
- [4] Shlyakhter I, Sabeti PC, Schaffner SF. C_{osi}2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 2014, 30(23): 3427–3429. [DOI]
- [5] Servedio MR. The evolution of premating isolation: local adaptation and natural and sexual selection against hybrids. *Evolution*, 2004, 58(5): 913–924. [DOI]
- [6] Daleszczyk K, Bunevich AN. Population viability analysis of European bison populations in Polish and Belarusian parts of Białowieża Forest with and without gene exchange. *Biol Conserv*, 2009, 142(12): 3068–3075. [DOI]
- [7] Alves DA, Imperatriz-Fonseca VL, Franco TM, Santos-Filho PS, Billen J, Wenseleers T. Successful maintenance of a stingless bee population despite a severe genetic bottleneck. *Conserv Genet*, 2011, 12(3): 647–658. [DOI]
- [8] Fu YX, Li WH. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol*, 1997, 14(2): 195–199. [DOI]
- [9] Li HP, Stephan W. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet*, 2006, 2(10): e166. [DOI]
- [10] Beaumont MA, Zhang WY, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*,

- 2002, 162(4): 2025–2035. [DOI]
- [11] Li HP. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol*, 2011, 28(1): 365–375. [DOI]
- [12] Lin K, Futschik A, Li HP. A fast estimate for the population recombination rate based on regression. *Genetics*, 2013, 194(2): 473–484. [DOI]
- [13] Gao F, Ming C, Hu WJ, Li HP. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3 (Bethesda)*, 2016, 6(6): 1563–1571. [DOI]
- [14] Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, 2008, 3(10): e3395. [DOI]
- [15] Huang YZ. The application of computer simulation in teaching population genetics. *Hereditas (Beijing)*, 1998, 20(4): 26–27.
黄远樟. 计算机模拟在群体遗传教学中的应用. *遗传*, 1998, 20(4): 26–27. [DOI]
- [16] Gao J, Pan SY, Cao J. Design and application of computer-assisted software for teaching and research of population genetics. *Hereditas (Beijing)*, 2008, 30(5): 642–648.
高婧, 潘沈元, 曹静. 群体遗传学教学与研究辅助软件的设计与应用. *遗传*, 2008, 30(5): 642–648. [DOI]
- [17] Sved JA. Genetics computer teaching simulation programs: promise and problems. *Genetics*, 2010, 185(4): 1537–1540. [DOI]
- [18] Vähä JP, Primmer CR. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Mol Ecol*, 2006, 15(1): 63–72. [DOI]
- [19] Ryman N, Palm S. POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Mol Ecol Notes*, 2006, 6(3): 600–602. [DOI]
- [20] Peng B, Amos CI. Forward-time simulation of realistic samples for genome-wide association studies. *BMC Bioinformatics*, 2010, 11: 442. [DOI]
- [21] Vonholdt BM, Stahler DR, Smith DW, Earl DA, Pollinger JP, Wayne RK. The genealogy and genetic viability of reintroduced Yellowstone grey wolves. *Mol Ecol*, 2008, 17(1): 252–274. [DOI]
- [22] Peng B, Kimmel M. Simulations provide support for the common disease-common variant hypothesis. *Genetics*, 2007, 175(2): 763–776. [DOI]
- [23] Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol*, 2010, 8(1): e1000294. [DOI]
- [24] Carvajal-Rodríguez A. Simulation of genomes: a review. *Curr Genomics*, 2008, 9(3): 155–159. [DOI]
- [25] Carvajal-Rodríguez A. Simulation of genes and genomes forward in time. *Curr Genomics*, 2010, 11(1): 58–61. [DOI]
- [26] Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet*, 2012, 13(2): 110–122. [DOI]
- [27] Yuan XG, Miller DJ, Zhang JY, Herrington D, Wang Y. An overview of population genetic data simulation. *J Comput Biol*, 2012, 19(1): 42–54. [DOI]
- [28] Kimura M. Evolutionary rate at the molecular level. *Nature*, 1968, 217(5129): 624–626. [DOI]
- [29] Ewens WJ. Mathematical population genetics. Berlin: Springer, 1979. [DOI]
- [30] Kingman JFC. The coalescent. *Stoch Proc Appl*, 1982, 13(3): 235–248. [DOI]
- [31] Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 1969, 61(4): 893–903. [DOI]
- [32] Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 1985, 111(1): 147–164. [DOI]
- [33] Krone SM, Neuhauser C. Ancestral processes with selection. *Theor Popul Biol*, 1997, 51(3): 210–237. [DOI]
- [34] Fu YX, Li WH. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor Popul Biol*, 1999, 56(1): 1–10. [DOI]
- [35] Wiuf C, Hein J. The coalescent with gene conversion. *Genetics*, 2000, 155(1): 451–462. [DOI]
- [36] Zöllner S, Pritchard JK. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 2005, 169(2): 1071–1092. [DOI]
- [37] Balloux F. EASYPOP (version 1.7): a computer program for population genetics simulations. *J Hered*, 2001, 92(3): 301–302. [DOI]
- [38] Wright S. Isolation by distance. *Genetics*, 1943, 28(2): 114–138. [DOI]
- [39] Strand AE. METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. *Mol Ecol Notes*, 2002, 2(3): 373–376. [DOI]
- [40] Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 2005, 21(18): 3686–3687. [DOI]
- [41] Peng B, Amos CI. Forward-time simulations of non-random

- mating populations using simuPOP. *Bioinformatics*, 2008, 24(11): 1408–1409. [DOI]
- [42] Peng B, Amos CI, Kimmel M. Forward-time simulations of human populations with complex diseases. *PLoS Genet*, 2007, 3(3): e47. [DOI]
- [43] Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 2008, 9: 364. [DOI]
- [44] Kim Y, Wiehe T. Simulation of DNA sequence evolution under models of recent directional selection. *Brief Bioinform*, 2009, 10(1): 84–96. [DOI]
- [45] Guillaume F, Rougemont J. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 2006, 22(20): 2556–2557. [DOI]
- [46] Neuenschwander S, Hospital F, Guillaume F, Goudet J. quantiNemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics*, 2008, 24(13): 1552–1553. [DOI]
- [47] Lambert BW, Terwilliger JD, Weiss KM. *ForSim*: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics*, 2008, 24(16): 1821–1822. [DOI]
- [48] Edwards TL, Bush WS, Turner SD, Dudek SM, Torstenson ES, Schmidt M, Martin E, Ritchie MD. Generating linkage disequilibrium patterns in data simulations using genomeSIMLA. In: Marchiori E, Moore J H, eds. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Berlin Heidelberg: Springer, 2008, 4973: 24–35. [DOI]
- [49] Hernandez RD. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 2008, 24(23): 2786–2787. [DOI]
- [50] Carvajal-Rodríguez A. GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics*, 2008, 9: 223. [DOI]
- [51] Padhukasahasram B, Marjoram P, Wall JD, Bustamante CD, Nordborg M. Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, 2008, 178(4): 2417–2427. [DOI]
- [52] Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*, 2007, 8(8): 610–618. [DOI]
- [53] Aberer AJ, Stamatakis A. Rapid forward-in-time simulation at the chromosome and genome level. *BMC Bioinformatics*, 2013, 14: 216. [DOI]
- [54] Sanford J, Baumgardner J, Brewer W, Gibson P, Remine W. Mendel's Accountant: a biologically realistic forward-time population genetics program. *SCPE*, 2007, 8(2): 147–165. [DOI]
- [55] Coombs JA, Letcher BH, Nislow KH. Pedagog: software for simulating eco-evolutionary population dynamics. *Mol Ecol Resour*, 2010, 10(3): 558–563. [DOI]
- [56] Carroll C, Fredrickson RJ, Lacy RC. Developing metapopulation connectivity criteria from genetic and habitat data to recover the endangered Mexican wolf. *Conserv Biol*, 2014, 28(1): 76–86. [DOI]
- [57] Laval G, Excoffier L. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, 2004, 20(15): 2485–2487. [DOI]
- [58] Excoffier L, Foll M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 2011, 27(9): 1332–1334. [DOI]
- [59] Yang T, Deng HW, Niu TH. Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. *BMC Bioinformatics*, 2014, 15: 3. [DOI]
- [60] Wiuf C, Hein J. Recombination as a point process along sequences. *Theor Popul Biol*, 1999, 55(3): 248–259. [DOI]
- [61] McVean GAT, Cardin NJ. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 2005, 360(1459): 1387–1393. [DOI]
- [62] Eriksson A, Mahjani B, Mehlig B. Sequential Markov coalescent algorithms for population models with demographic structure. *Theor Popul Biol*, 2009, 76(2): 84–91. [DOI]
- [63] Marjoram P, Wall JD. Fast “coalescent” simulation. *BMC Genet*, 2006, 7: 16. [DOI]
- [64] Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data. *Genome Res*, 2009, 19(1): 136–142. [DOI]
- [65] Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 2002, 18(2): 337–338. [DOI]
- [66] Hellenthal G, Stephens M. msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics*, 2007, 23(4): 520–521. [DOI]
- [67] Ramos-Onsins SE, Mitchell-Olds T. Mlcoalsim: multi-locus coalescent simulations. *Evol Bioinform*, 2007, 3: 41–44. [DOI]
- [68] Liang LM, Zöllner S, Abecasis GR. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 2007, 23(12): 1565–1567. [DOI]

- [69] Currat M, Ray N, Excoffier L. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol Notes*, 2004, 4(1): 139–142. [DOI]
- [70] Mailund T, Schierup MH, Pedersen CNS, Mechlenborg PJM, Madsen JN, Schausser L. CoaSim: a flexible environment for simulating genetic data under coalescent models. *BMC Bioinformatics*, 2005, 6: 252. [DOI]
- [71] Antao T, Beja-Pereira A, Luikart G. MODELER4SIMCOAL2: a user-friendly, extensible modeler of demography and linked loci for coalescent simulations. *Bioinformatics*, 2007, 23(14): 1848–1850. [DOI]
- [72] Anderson CN, Ramakrishnan U, Chan YL, Hadly EA. Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, 2005, 21(8): 1733–1734. [DOI]
- [73] Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*, 2013, 9(10): e1003905. [DOI]
- [74] Spencer CCA, Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 2004, 20(18): 3673–3675. [DOI]
- [75] Teshima KM, Innan H. mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, 2009, 10: 166. [DOI]
- [76] Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 2010, 26(16): 2064–2065. [DOI]
- [77] Ilves KL, Huang W, Wares JP, Hickerson MJ. Colonization and/or mitochondrial selective sweeps across the North Atlantic intertidal assemblage revealed by multi-taxa approximate Bayesian computation. *Mol Ecol*, 2010, 19(20): 4505–4519. [DOI]
- [78] Wernsdörfer H, Caron H, Gerber S, Cornu G, Rossi V, Mortier F, Gourellet-Fleury S. Relationships between demography and gene flow and their importance for the conservation of tree populations in tropical forests under selective felling regimes. *Conserv Genet*, 2011, 12(1): 15–29. [DOI]
- [79] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. *Nat Rev Genet*, 2010, 11(9): 647–657. [DOI]
- [80] Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*, 2010, 10(3): 564–567. [DOI]
- [81] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*, 2013, 30(12): 2725–2729. [DOI]
- [82] Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*, 1997, 13(3): 235–238. [DOI]

(责任编辑: 赵方庆)