深圳中学研究性学习结题报告

2020 年 12 月 6 日

课题名称 用信息方法研究遗传学问题

课题负责人 杨景云

课题成员
- 杨景云 2019530093
- 杨培源 2019530256
- 孟涵宇 2019530676
- 靳柏舟 2019530303
- 张杜宇 2019530484
- 张佳皓 2019530414
- 国星 2019530568
- 杨广源 2019530743
- 杨邓弘 2019530307

指导教师 李丽华老师

所在班级 高二 (9) 班

摘要

本小组通过信息学的方法，以一种全新的视角来解决传统的生物遗传学问题，并以此总结了一种通用的方法来帮助高中生更好理解并解答高中的遗传学问题。本文运用了建立数学模型的方法，并将本组的研究成果总结成论文。论文中本组通过建立表现型比例、基因型比例等数学模型，提出了高效求解生物遗传学问题的方法。论文内容摘要如下：

1. 论文先对真值运算符这一概念进行了二进制的约定，并定义了本组建立的数学模型中的基因集合、基因片段、生成函数、表现型映射等关键概念，为后文的生成函数运算和不同情况问题的解答奠定了基础；
2. 在定义了基础概念后，论文提出了最基础的只有显隐性的群体自由交配后子代基因型比例问题的求法，并将其分为配子生成函数的求法、基因片段生成函数的求法这两个部分，对实现问题的解答所需的朴素做法和快速做法的具体方法进行了讲解。在配子生成函数求解中，论文提出了针对等位基因数目多少的不同情况提出了求解配子出现次数函数及运用 Trie 树求解这两种不同的方法；在基因片段生成函数求解中，论文则运用将该函数转换成集合生成函数，并使用其快速算法集合并卷积的 FWT 算法（快速莫比乌斯算法）对此类问题进行了较为简洁的解答，总结出了便于高中生解决此类问题的简洁纯笔算方法，并建立了解决此类问题的代码。
3. 论文对于上述的配子及基因片段生成函数求法针对时间复杂度（即运算所需运算时间）这一角度进行了进一步的优化，并讲解了 FWT 算法的一些性质以及一些特殊情况时使用的矩阵运算方法，且附上了多维广义离散傅里叶变换所需的代码，方便他人进行运算。
4. 论文针对互补基因、累加作用、上位效应和抑制作用这四种较为常见的现象专门进行了介绍并设计了运算方法，并分类讨论了不允许自交的情况和存在多对不同作用等位基因时的优化方法。其中对于纯合致死问题这一典型问题进行了探讨，并运用子集卷积的方法提出了朴素做法及快速做法这两种在不同情况下各有优势的做法。

关键词：信息学方法、遗传学问题、数学模型、二进制约定、生成函数、表现型映射、Trie 树、FWT 算法、时间复杂度、子集卷积。

1 引言

高中生物中的遗传学问题一直是学习中的难点。由于缺少既高效又通用的方法，很容易在复杂的表现型比例, 基因型比例的计算中出错。同时, 在设计杂交方案等具有综合性的问题中也会遇到困难。而本文试图通过引入如生成函数等更高级的数学工具，以及计算机方法来探索新的求解方法。

我们提出的数学建模思路来源于一个出现在教辅书上的经典问题 [1]: 基因型为 AaBB 的个体自交，但含有 a 基因的个体有 $\frac{1}{2}$ 的几率不能存活。给出的解法 [2]: 含有 A 基因的配子的概率为 $\frac{2}{3}$，含有 a 基因的配子的概率为 $\frac{1}{3}$。如果把不同配子看成多项式的系数，基因型看做指数，那么杂交过程就可以看成多项式乘法。$(\frac{2}{3}x^{\mathtt{A}} + \frac{1}{3}x^{\mathtt{a}})(\frac{1}{2}x^{\mathtt{B}} + \frac{1}{2}x^{\mathtt{B}}) = \frac{2}{3}x^{\mathtt{AB}} + \frac{1}{3}x^{\mathtt{aB}}$，那么两种基因型的比例就是 $2:1$。这种方法将生物学问题转化成纯粹的数学问题, 但引入算式的部分缺乏严谨性。本文将运用这种方法的思想，将这种算式抽象化为集合幂级数。通过将 AB 这样的基因集合定义为广义的指数, 将模型在数学上严格化，进而得到通用的手工求解做法。

更重要的是, 随着基因片段长度的增加, 这种方法计算的复杂度会大大增加。这时我们就可以引入计算机手段来求解该问题。如果直接模拟多项式的乘法过程，算法时间复杂度依然很高。此时就需要运用能求解集合幂级数卷积的高效算法：快速沃尔什变换 (Fast Walsh Transform, FWT) 和快速莫比乌斯变换 (Fast Mobius Transform, FMT)。

2 文献综述

19 世纪末，遗传学的基本定律已经由孟德尔 (Gregor Johann Mendel)，摩尔根 (Thomas Hunt Morgan) 等人提出，并在细胞学研究中证明。基因的分离定律 (Law of Segregation) 和自由组合定律 (Free Combination Law of Gene Independent Assortment) 使得遗传学中的出现频率问题可以用组合数学计算。[1]

作为组合数学的高效工具，生成函数 (Generating Function, 又称母函数) 最初由棣莫弗 (Abraham De Moivre) 提出, 最初是用于求线性递推数列的通项公式。[2] 生成函数将数列的下标作为指数，下标对应的值作为系数, 就可以把数列问题转化为代数上的形式幂级数运算。在求解组合数学问题中, 只需把数列定义为组合问题在不同规模下的方案数即可。[3]。将集合定义为广义的指数，就可以得到集合幂级数, 可以求解下标为集合的数列，进而求解和集合有关的组合数学问题。

这样，我们已经有了一套成熟的数列理论来求解相关的组合数学问题。但是，要使用计算机来处理幂级数的运算，需要更高效的算法。1965 年 James Cooley 与 John Tukey 提出的快速傅里叶变换 (Fast Fourier Transform, FFT) [4] 可以在 $\mathcal{O}(n \log n)$ 的时间复杂度内处理下标为正整数的多项式卷积。1976 年出现了可以求解集合为下标的算法（子集卷积）, 即快速沃尔什变换 (Fast Walsh Transform, FWT) [5]。沃尔什变换利用分治的思想和 Hadamard 矩阵加速了求解过程 [6]。2007 年 Andreas Björklund 总结了前人的工作, 用 Möbius 变换和反演计算在任意环中进行加法和乘法的子集卷积，得到了 $\mathcal{O}(m^2 2^m)$ 的子集卷积，对 $\mathcal{O}(3^m)$ 的传统算法进行了改进。具体来说，如果输入函数的整数范围为 $[-M, M] \cap \mathbb{Z}$，则它们的子集卷积可以用 $\mathcal{O}(2^m \log M)$ 时间求解。还利用矩阵解决了高维子集卷积问题 [7]. 这些算法已经足够我们进行基因相关的组合计数。

3 研究方法

我们的研究从基础的遗传学计算问题展开，运用建立数学模型的方法表达生物学中例如基因型、配子、自由组合、表现型和随机结合等概念。由此，面对特定的遗传学问题的时候，我们就可以利用我们已知的函数的性质去研究，然后通过计算机计算的方法快速地给出结果，从而使运用这种工具的人能够通过按一下鼠标的方式了解计算结果，从而对不同基因型的组合产生的效果有一个大概的认知。

我接下来将介绍研究中使用的计算机学的基础工具：

1. 我们定义了基因集合的概念，用一个从该集合向自然数集的映射给这些集合标号，用一个 n 维的向量来表达 n 对不同的等位基因进行复合的结果，对于个体而言，向量的每一“维”是一个二元自然数对；对于配子而言，每一“维”是一个自然数。定义运算 来表达两个配子之组合。

2. 利用生成函数的概念来表示“基因生成函数”，运用这个函数来模拟个体产生配子和两个个体之杂交。

3. 利用表现型映射来刻画基因之间的关系（完全显性、不完全显性、隐性或如喷瓜那种更复杂的显隐关系）、利用另一个 n 维向量的形式表达这种表现性。

4. 通过被称为卷积的概念来计算具体的 n 维向量间的关系。

此后，对于解决特定的问题，我们会使用一些特定的信息学技巧来简化计算量。这部分体现的信息学较多，为保证该模块的篇幅不过于长，我们会在随后提供的 pdf 中进行详细的说明。

对于我们目前的努力，优点和不足如下：

1. 我们任务目标是让高中生能够更好地理解遗传学，我们采取的思路是让他们提供基因的组合，通过感受基因组合的结果，对遗传学的一些规律有基本的认识。对于这个思路，我们采用信息学的方法能够很快地给出精确的结果，因此这个思路方面我们无疑是成功的。

2. 我们对遗传学的问题考虑得较为周到，这使得我们能够给出高中生绝大部分关于遗传学的问题的答案。

3. 我们已经给出了相关的网站，因此我们的研究目标已经完成。

4. 我们的研究尚未收到反馈，采取让同学感受基因组合结果的思路是否能够真正更好地让同学对遗传问题的结论有更好的认识还是未知数，这也是下一步我们研究的主要目标。

4 结果分析与讨论

5 结论

如附件论文所示，通过以信息学方式考察遗传问题，可以较正常解法更快速地得到结果。通过卷积与快速傅里叶变换等数学与信息技巧，我们找到了用于处理不同条件与情况下遗传

学问题的公式化解法，并以此制造了一个网站 https://github.com/Shenzhen-Middle-School-OI-team/Power-Series-Calculation-Model-of-Genome-Combination。这个网站主要针对学习高中遗传学的学生，使用者可以便捷地研究不同的遗传学问题，方便学生理解遗传的概念。

本研究的成果体现了跨学科交互的作用：通过将生物学问题引入信息学处理方式，并辅以数学的计算技巧，才得到了最终的研究结论。这个研究过程不仅加深了我们对各科知识的掌握，也启发了我们对于问题要以多种思路分析的思维方式。

参考文献

[1] 葛明德吴相钰, 陈守良. 陈阅增普通生物学. 高等教育出版社, 2009.

[2] Donald E Knuth. *The Art of Computer Programming, Volume 1, Fascicle 1: MMIX–A RISC Computer for the New Millennium*. Addison-Wesley Professional, 2005.

[3] Ronald L Graham, Donald E Knuth, Oren Patashnik, and Stanley Liu. Concrete mathematics: a foundation for computer science. *Computers in Physics*, 3(5):106–107, 1989.

[4] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[5] David K Maslen and Daniel N Rockmore. Generalized ffts—a survey of some recent results. In *Groups and Computation II*, volume 28, pages 183–287. American Mathematical Soc., 1997.

[6] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6, 1987.

[7] Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. Fourier meets möbius: Fast subset convolution. STOC '07, page 67–74, New York, NY, USA, 2007. Association for Computing Machinery.