QUNTIYICHUAN DE
XINXIMOXING FENXI

群体遗传的信息模型分析

张宏礼　刘胜军　著

東北林業大學出版社

群体遗传的信息模型分析

张宏礼　刘胜军　著

东北林业大学出版社

图书在版编目（CIP）数据

群体遗传的信息模型分析/张宏礼，刘胜军著. —哈尔滨：东北林业大学出版社，2009.7

ISBN 978-7-81131-484-7

Ⅰ.群… Ⅱ.①张…②刘… Ⅲ.群体遗传学-研究 Ⅳ.Q347

中国版本图书馆 CIP 数据核字（2009）第 115009 号

责任编辑：倪乃华

封面设计：彭 宇

NEFUP

群体遗传的信息模型分析

Quntiyichuan De Xinximoxing Fenxi

张宏礼 刘胜军 著

东北林业大学出版社出版发行

（哈尔滨市和兴路 26 号）

哈尔滨天兴速达印务有限责任公司印装

开本 787 × 960 1/16 印张 12 字数 215 千字

2009 年 7 月 第 1 版 2009 年 7 月 第 1 次 印 刷

印数 1—1 000 册

ISBN 978-7-81131-484-7

定价：25.60 元

本书由黑龙江省教育厅科学技术研究项目“群体遗传的信息模型分析”资助，项目编号：11531255。

前　言

遗传学是生物科学的基础，是生命科学领域中发展最为迅速的前沿学科之一，而群体遗传学作为现代生物学分支在中国的发展曾经经历了一个曲折的过程，目前从事这一领域研究的人员不多。但是，随着人类社会飞速发展中带来的许多问题，为了追求人类和环境的和谐共存、人类的可持续发展，这一领域的研究越来越受到重视。

群体遗传学研究生物群体的遗传结构、遗传结构的变化规律以及种群演化规律，群体遗传学的发展和生物进化理论密切相关。数学模型的建立、分析和验证是群体遗传学中不可或缺的研究手段。数学模型是对有关生物系统特性的高度概括，可以发现生物学家未曾预料到的在进化中起作用的力。近年来，一些学者应用信息论方法进行了有关群体遗传学的研究。一方面，取得了一些和传统的数学方法一致的结论；另一方面，也获得了一些对新成果、新的信息学的解释。这一数学方法的好处是：可以避免由于对等位基因赋值导致的不合理性，并容易将所得结果推广到更一般情形。

本书希望对近年来应用信息论方法研究群体遗传学的成果做一个比较系统的总结，主要内容包括随机交配下群体遗传的信息模型、近亲交配下群体遗传的信息模型、选型交配下群体遗传的信息模型、性连锁群体的信息模型、突变和选择下群体遗传的信息模型、亲属关系和相似性度量的信息模型。

本书的第2,3,6,7,8章由黑龙江八一农垦大学文理学院的张宏礼博士撰写，约11万字；第1,4,5章由黑龙江八一农垦大学动物科技学院的刘胜军博士撰写，约10万字。最后由张宏礼博士统稿全书。

西北农林科技大学的袁志发先生对本书进行了审阅，提出了许多宝贵的意见，在此深表谢意。

由于作者学识有限，不足之处请读者批评指正。

著者

2009年4月

目　录

1 绪 论

1.1 群体遗传学的发展

群体遗传学(Population genetics)是研究生物群体的遗传结构、遗传结构的变化规律以及种群演化规律的遗传学分支。群体遗传学的发展和生物进化理论密切相关。生物进化是生物群体的遗传组成部分或全部的不可逆的一系列转变,这种转变基本上是基于生物与其环境相互作用的改变。最初,人类并未认识到遗传学和进化论有什么必然的联系。随着对基因认识的不断深入,遗传学和进化论找到了共同的渊源。从原始生命开始,基因担负着传递遗传信息的功能。通过基因突变、基因重组等途径,使遗传信息多样化和复杂化,在选择和隔离等机制的作用下,生物趋异形成多个种。在基因基础上群体遗传学和进化论实现了完美而和谐的统一。

有关群体遗传学的起源可以追溯到英国博物学家达尔文(C. Darwin, 1809—1882)提出的生物进化的科学理论。达尔文于1859年发表了《物种起源》一书,提出了关于生物进化的自然选择学说,强调微小变异的累积对于物种进化的重要性。达尔文的进化论学说可以概括为三原则:①变异原则:同一种群的个体间在生理、形态和行为等方面有差异;②遗传原则:有亲源关系的个体间的相似程度比无亲源关系的个体间的相似程度大;③选择原则:在一定环境下,群体内的一些类型比另一些类型具有更强的生存和繁殖能力。尽管达尔文接受泛生论和获得性状遗传的观点,但这也丝毫掩盖不了以三个原则为核心的进化的自然选择学说的光芒,至今仍为人们广泛接受。

在达尔文之后,进化论经历了两次重大的修正。魏斯曼(A. Weismann)首先向"泛生论"、"用尽废退"和"获得性状遗传"挑战,提出了"种质"学说,并最早对达尔文的进化论进行了改造。"种质"学说把细胞分为种质细胞系和体质细胞系,认为"体质"是由"种质"产生的,"种质"不灭,世代相继;环境只能影响"体质",而不能影响"种质",故"获得性状"不能遗传。经魏斯曼等人

改造后的达尔文进化论称为“新达尔文主义”。尽管魏斯曼的“种质”学说也存在缺点和错误,但其科学的合理内核对以后遗传学的发展有相当大的影响。

群体遗传学的基本原理是由奥地利人孟德尔(Mendel)最早揭示的。1856~1864年,孟德尔做了八年的豌豆杂交试验。结合前人的工作,孟德尔做出了遗传因子的分离和重组的假设,并于1865年在布隆自然科学协会的学术例会上相继报告了他的研究成果。1866年,孟德尔在布隆自然科学协会会刊第4卷上发表了他的论文《植物杂交实验》,但此后被学术界忽视达34年之久。直到1900年,荷兰的德弗里斯(De Vries)、德国的科伦斯(Correns)和奥地利的丘歇马克(Tschermak)三位科学家经过大量的植物杂交工作,在不同的地点、不同的植物上,取得了与孟德尔实验相同的结果,验证了孟德尔群体遗传学研究结果的正确性。这一年标志着整个遗传学的诞生,孟德尔是遗传学的奠基人。孟德尔遗传理论的核心是“颗粒遗传”学说,认为遗传因子是呈颗粒状的,互不融合,互不沾染,独立分离,自由组合。此后相当长的一段时间遗传学是按这一思路发展的。

1908年英国数学家哈迪(Hardy)和德国医生温伯格(Weinberg)独立发表了群体遗传平衡的文章,将孟德尔定律用于随机交配的大群体,正式提出平衡定律,现称为Hardy-Weinberg平衡定律,为群体遗传学的诞生奠定了第一块基石。此后不久,约翰森(Johanson)提出了著名的纯系学说,提出了基因型与表现型的概念,区分了遗传变异和非遗传变异。到20世纪20~30年代,费歇(Fisher)、郝尔登(Haldane)和拉伊特(Laright)等发表了《自然选择的遗传理论》、《进化的起因》以及《孟德尔群体内的进化》等著作,将达尔文的自然选择与孟德尔定律结合起来了,形成了群体遗传的基本理论体系。由于这些进展,引起对“自然选择”学说本身以及与其相关的概念(如适应概念、物种概念)做出修正,如用繁殖的相对优势来定义适应;用个体或基因型对后代基因库的相对贡献定义适应度;选择和适应不再是“生存斗争,适者生存”,而是繁殖或基因传递的相对差异等。经过这次修正的达尔文进化论,称为“现代综合论”。基于进化的现代综合论之下的达尔文三原则被认为是群体遗传学的同一语。1937年,杜布赞斯基(T. Dobzhansky)出版的《遗传学和物种起源》是群体遗传学的一个里程碑,标志着群体遗传学理论的基本骨架形成。1955年,华裔学者李景钧(C. C. Li)所著的《群体遗传学》总结了20世纪前半期群体遗传学的几乎全部成果,内容丰富,阐述精湛,标志着群体遗传学理论体系的系统化和完善化。

伴随着研究手段的不断改进,人类对遗传学的研究逐渐向分子水平迈进。分子遗传学可以追溯到1944年艾弗里(O. T. Avery)等人的肺炎球菌转化实

验证明了 DNA 是遗传转化因子。但是突破性的事件却是 1953 年 4 月 25 日出版的《自然杂志》刊登的华生(James D. Watson)和克里克(Francis Crick)的研究论文《核酸的分子结构——脱氧核糖核酸的结构》。该论文提出了 DNA 双螺旋结构模型,把艾弗里等人验证 DNA 是遗传物质的研究成果建立在更牢固的基础之上,开创了分子遗传学的新纪元,标志着遗传学以及生物学进入到分子水平的新时代,遗传学从此在各个方面取得了飞速发展。1977 年,桑格(F. Sanger)等人完成了噬菌体 ΦX174 的全部碱基序列测序,确立了 DNA 序列分析的新战略和新方法,从而使分子遗传学又进入一个崭新的时代。近 40 年来,分子遗传学取得了极其巨大的成就,已成为生命科学的带头学科之一,有利地促进了生命科学中各分支学科的发展。1997 年英国维尔莫特(Lan Wilmut)等人宣布体细胞克隆绵羊成功,从而突破了遗传学固有的理论体系,使世界为之一惊。随后,在世界范围内掀起了克隆热潮,先后有多种动物克隆成功。一些国家正在进行克隆人的研究,成功只是时间的问题。体细胞克隆的成功已超出遗传学的范畴,引起人类的广泛思考。

与分子遗传学的发展相适应,许多学者试图把生物大分子的结构变化与群体遗传学的理论结合起来,从分子水平阐明生物群体的系统进化理论。围绕分子水平上的遗传变异保持机制问题,形成了两种对立的学说:一种是以费歇(Fisher)、郝尔登(Haldane)、拉伊特(Laright)、莱特(Wright)、杜布赞斯基(Dobzhansky)等为代表的"现代综合论",认为自然选择仍起重要作用;另一种是以日本木村资生(Kimura)为代表,于 1968 年提出的分子进化的中性学说,认为分子水平上的变异几乎是中立的,与自然选择无关,其依据是:生物大分子的大部分进化变异是中性突变基因通过遗传漂变随机固定的结果。分子进化的中性学说和进化的现代综合论中的许多细节问题,有待于科研工作者在群体遗传和分子遗传领域的研究中进一步阐明。这些争论也预示着达尔文的进化理论即将面临着再次的修正。新的修正将涉及进化的速度、进化的方向、进化的动力等问题。之所以称之为修正是因为生物化学、分子生物学、分子遗传学等学科的研究还无法否定达尔文进化论的核心原则,即遗传的原则、变异的原则和选择的原则。为了解决上述争论,有必要从不同的角度、应用不同的方法更深入地研究群体遗传学。

1.2 群体遗传学中的数学方法

前面已经谈到,群体遗传学是在 20 世纪 20 年代和 30 年代由一些理论家

发展起来的。他们用严密的数学模型发现了群体内基因和基因型的动力学，扩展了孟德尔和达尔文的原理。因此，在某种程度上，群体遗传学一直是一门独特的学科。群体的遗传变异受多种因子的影响，进化中许多重要问题是无法通过对自然群体或实验群体的观察或实验来解决的。例如，综合论者曾经通过生物实验，证明了自然选择是进化的动力，但是要证明它所起作用的程度，它进行得多快，它的限制是什么，它能创造哪些遗传系统以及如何与其他进化力相互作用，进行数学处理是必要的。

数学模型是对有关生物系统特性的高度概括，可以发现生物学家未曾预料到的在进化中起作用的力。典型的例子是通过数学模型，提出了从有限的群体大小和由距离因素造成的群体隔离所导致的基因库分化的原理，即莱特(Wright)的随机漂变。虽然随机漂变在进化中的重要性仍然需要实验学家来证明，但它作为一个数学事实却是重要的，因此这个原理是从数学模型获得的理论认识。莱特(Wright, 1960)一直强调理论与观察的结合，他认为："数学理论是在个体水平和群体水平上发现实际知识的媒介，必须从个体水平的假设和从群体结构的模型中推论出在群体中将期望到什么，然后在任何与观察相矛盾的基础上来修正它的假设和数学模型。进化是在群体中发生的，没有数学理论将群体中的现象与个体现象联系起来，就不会有清楚的进化思想。"另外，数学模型是以某些简化的假设为前提的，这些假设的正确性必须用实验数据加以检验。

受分子遗传学飞速发展的影响，不少科技工作者改变自己的研究方向，由统计遗传学领域转向分子遗传学领域，甚至有人认为统计遗传学应该让位于分子遗传学。但是，近些年来遗传学的发展实践证明，事实远非如此。仅仅依靠分子遗传学而不利用合适的数学方法，许多研究耗时费力，尤其是随着基因组和后基因组时代的到来，数学方法越来越显现出重要性，只有把二者结合起来，而不是顾此失彼，才是科学发展的正确方向。

长期以来，群体遗传学中所用的数学模型基本上都是传统的统计学模型。群体遗传学所研究的世代传递过程本身也是一个信息传递的过程，因此信息论模型也应该是研究该门学科的一种数学模型，而不应仅限于以往的统计学模型。本书正是作者在近十年来的研究基础之上，总结我们自己和国内其他学者应用信息论方法研究群体遗传学问题的成果而完成的，以此丰富群体遗传学的内容。

2 应用信息论方法研究群体遗传学的基础

信息论从诞生的时刻起就引起了众多学者的注意，他们竞相应用信息论的方法去理解和解决本领域中的问题，当然，生命科学也不例外。这是因为生命体是一个复杂的系统，生命过程是一个高度复杂的过程。生命有机体需要物质和能量支持，所以生命体也是一个信息系统。从发展的观点看，把生命过程看做是一个高级的信息过程，将有助于理解生命的本质，有利于推动和促进生命科学的发展。

2.1 信息论的形成和发展

2.1.1 信息

信息一词在我国由来已久。辞海中曾记有“梦断美人沉信息，目穿长路依楼台”的诗句，可见信息泛指音讯和消息。在近代，信息一词又被用作英语中 information 的译名，information 在牛津英文字典里给出的解释是“某人被通知或告知的内容、情报、消息”。在这样的解释中，信息一词显然不是作为科学名词或技术术语来定义的，因此无法作更深入的推敲。这种目前尚难定义的含义模糊、难于捉摸的信息我们可以称它为广义理解的信息。

随着计算机的发展，无论是在计算机界还是在工业界都希望有一个名称能把所有处理对象统统包含在内。信息这一名称恰好符合这一要求，因为只有这样一个含糊术语才能对多种多样且在不断涌现的对象得到一个统一的、全面的、不需时时改变的表达，从而信息作为一个技术术语而广泛出现。作为技术术语的信息其意义要比广义信息含义具体得多，但是仍然是比较笼统和含糊不清的。

信息作为一个可以用严格的数学公式定义的科学名词首先出现在统计数学中，随后又出现在通信技术中。无论是在统计数学中还是在通信技术中定

义的信息都是一种统计意义上的信息,我们简称为统计信息。统计信息是非常明确的,同时其适用范围要比广义信息狭隘得多,我们在本书中所用的信息正是关于统计信息的理论。

统计信息是一个抽象然而明确的概念,是一种有明确定义的科学名词,它与内容无关,而且不随信息具体表达形式的变化而变化,因而独立于形式。它反映了信息表达形式中统计方面的性质,是统计学上的抽象概念。

2.1.2 信息论的形成和发展

从历史上看,信息论的形成是两部分人共同努力的结果,一部分是通信工程方面的学者,另一部分是统计学家。这两部分人虽然研究的是同一领域的问题,但是他们感兴趣的方面和侧重点是有差异的。这种情况从信息论产生时起一直保持到现在,今天从事信息论研究工作的人仍然由这两部分人组成。

2.1.2.1 通信技术的理论基础

信息论的形成与发展最主要是以通信技术基础理论的形式逐步形成和发展起来的。这一点有它内在的原因。一方面,广义信息的含义极其复杂,而通信本身只涉及信息的表现形式或者说仅对信息的表现形式感兴趣,而这是广义信息最简单最基础的方面,因此,我们可以认为正是从最简单的方面取得了突破,才形成了信息论。另一方面,当通信技术得到广泛发展和应用以致形成通信网以后,人们自然要问:既然交通解决物质的运输,电网解决能量的传输,那么通信传递的究竟是什么?而信息论正是对这一问题的全面和系统的回答。但是不要以为人们只要想到"通信传送的究竟是什么?"就会自然地导致信息论的诞生。只有当人们无法实现"准确再现"时,理论上的追根究底也才有了动力,并导致信息论的诞生。

20 世纪 30 年代以前通信的主要目标还集中在如何使发送信号无失真地送到接收端,所用的分析方法还是分析确定性信号的方法。虽然 Hartley 定义了信息量,但是还不是一个统计的概念,因此其意义还是相当有限的。20 世纪 30 年代,由于通信技术的提高以及随后的第二次世界大战的爆发,使通信中的噪声和抗干扰问题逐渐突出,抗干扰的通信方法先后出现。进入 20 世纪 40 年代以后,通信的理论已经全面走上统计分析的道路,抗干扰已经取代抗失真成为通信研究中的中心问题。在这样的背景下,1948 年在申农(Shannon)的《通信的数学理论》和维纳(N. Wiener)的《控制论》中,两人几乎同时提出了信息的统计定义,这两个文献后来被认为是信息论的经典著作。在 1948 年以后的十余年中,申农对信息论的发展作出了巨大的贡献。在 1973

年出版的信息论经典论文中,申农是49篇论文中12篇论文的作者。迄今为止,信息论的主要概念除了通用编码外几乎都是申农首先提出的。此外,申农还证明了一系列的编码定理。这些编码定理不但给出了某些性能的理论极限,而且也是对申农所给基本概念的重大价值的证明。由于申农的这一系列的贡献,申农被认为是信息论的创始人。20世纪50年代起,通信技术界就把主要的精力转向信源编码和信道编码的具体构造方法上。几十年来,在无失真信源编码方面、有失真信源编码方面、面向数字信道的信道编码方面都取得了稳步的进展。在半个世纪的历程中,信息论作为通信技术基础理论的意义已经有了重大的发展。

2.1.2.2 统计数学的一个分支

从历史上看,信息作为一个科学名词首先出现在统计数学中。1925年,即R. V. Hartley发表信息量定义的前三年,统计学家R. A. Fisher就从古典统计理论的角度定义了一种信息量,现在一般称为Fisher信息量。

申农的论文《通信的数学理论》发表以后,一方面由于文中出现的"信息"一词引起各相关领域的重视,同时由于文中所涉及的数学问题而引起统计学家的兴趣。数学家们纷纷把申农的基本概念和编码定理推广到更一般的信源模型、更一般的编码结构和性能量度,并给出严格的证明。在发展信息论的概念方面,前苏联数学家A. N. Kolmogorov有突出的贡献。1956年他提出信息量的一般定义,1958年他指出熵相等是动力系统同构的必要条件,开辟了遍历理论的新方向。1968年他又提出定义信息量的三种途径,这一工作后来得到G. J. Chaitin的发展,并在1987年建立了算法信息的理论。E. T. Jaynes在1957年提出了最大熵原理的理论。1959年S. K. Kullback系统地论述了鉴别信息的概念、定义及其和Fisher信息量、Shannon信息量的关系,J. S. Shore等人在1980年后发展了该理论。这些成果大大丰富了信息理论的概念、方法和应用范围,把信息的统计定义进一步推广并对非统计意义的信息给出了一种量度,而且信息量度的意义也不再限于信源编码和信道编码领域。

2.2 信息论的基本概念与理论

信息论诞生后的几十年来,对信息论的研究取得了很多进展,但是迄今为止实际问题中应用最广泛的概念主要有三个:熵(entropy)、互信息(mutual information)和鉴别信息(discrimination information)。本书中主要应用前两个概念,这两个概念都是申农在1948年发表的论文《通信的数学理论》提出的。

2.2.1 Shannon 信息熵

信息论创始人申农在1948年发表的论文《通信的数学理论》标志着信息论的诞生。习惯上,以申农理论为核心的经典信息理论又称狭义信息论或统计信息论。

申农指出,存在这样的不确定性度量,它是关于概率分布 $p_1,p_2,\cdots,p_N$ 的函数 $f(p_1,p_2,\cdots,p_N)$,并且满足如下3个先验条件。

(1)连续性条件:$f(p_1,p_2,\cdots,p_N)$ 应是 $p_n(n=1,2,\cdots,N)$ 的连续函数;

(2)等概率时为单调函数:$f(1/N,1/N,\cdots,1/N)=g(N)$ 应是 N 的增函数;

(3)可加性条件:当随机变量的取值不是通过一次试验而是通过若干次试验才最后得到时,随机变量在各次试验中的不确定程度应该可加,且其和始终与通过一次试验取得结果的不确定程度相同,即

$$f(p_1,p_2,\cdots,p_N)=f[(p_1+p_2+\cdots+p_K),p_{K+1},\cdots,p_N]+(p_1+p_2+\cdots+p_K)f(p'_1,p'_2,\cdots,p'_K)$$

其中

$$p'_k=\frac{p_k}{(p_1+p_2+\cdots+p_k)},k=1,2,\cdots,K$$

申农的研究证明,满足上述3个条件 $f(p_1,p_2,\cdots,p_N)$ 的形式可唯一地表示为

$$f(p_1,p_2,\cdots,p_N)=-C\sum_{n=1}^{N}p_n\log p_n$$

其中 C 为大于零的常数。申农将其作为随机变量不确定的量度,今天称之为Shannon信息熵(简称信息熵或熵),并记为

$$S(p_1,p_2,\cdots,p_N)=-\sum_{n=0}^{N}p_n\log p_n$$

当有多个随机变量时,为区别不同随机变量的熵,可将熵写成 $S(X)$,$S(Y)$,以分别表示随机变量 X 或 Y 的熵。熵函数中的常数取不同的值也会有不同的意义,这在本书以后的内容中会看到。

2.2.2 Shannon 信息熵可以作为信息的量度

Shannon信息熵之所以可以作为信息的量度,是因为对于随机变量而言,其取值是不确定的。在随机试验之前,我们只知道了所有可能取值。试验后,