

文章编号: 1005-0906(2013)01-0001-08

现代数字育种技术的研究进展

董春水¹, 才卓²(1《世界种业》网, [Http://the.worldseeds.com](http://the.worldseeds.com); 2《玉米科学》编辑部, 长春 130033)

摘要: 现代计算机技术促进了作物育种技术的数字化进程。随着作物性状数据采集技术的发展成熟, 作物育种过程中产生的数据呈指数增长, 促使传统作物育种技术变革。作物育种过程中产生的数据类型十分复杂, 数据的存储、分析和利用是作物育种技术的一个组成部分, 促成了现代数字育种技术的迅猛发展。通过分析玉米育种过程中涉及到的各类数据, 提出概念性的作物育种数据管理系统框架, 并阐明各类数据之间的相互关系。完善的育种数据管理系统除核心数据库外, 还包括多项数据分析模块, 包括系谱树和亲缘分析、分子标记和基因定位、数据采集和性状分解、杂交组合和后代选择、育种策略分析、田间试验设计和统计分析、生长发育系统模拟和基因功能和调控网络等。数据分析模块可以根据作物育种实践的实际需要进行组合, 但并非所有模块都是必需的。

关键词: 玉米育种; 数据库管理; 数字育种技术; 数据分析; 概念框架

中图分类号: S513.035

文献标识码: A

Advances in Modern Data-driven Breeding Technologies

DONG Chun-shui¹, CAI Zhuo²(1. *The Net Station of World Seeds*;2. *Journal of Maize Sciences*, Changchun 130033, China)

Abstract: The modern computer technologies have greatly fostered the modern data-driven crop breeding technology. As data acquisition technologies getting mature, data generated in the crop breeding process grow exponentially and explosively. Traditional crop breeding technology has been revolutionized to accommodate the massive data in breeding processes. Because the data types generated in crop breeding processes are extremely complex, storing, analyzing, and effectively utilizing them become increasingly challenging and directly lead to the rapid development of modern data-driven breeding techniques. A conceptual framework of crop breeding data management system was provided based on the thorough analysis of the diverse types of data generated in maize breeding programs, and the relationship were also enucleated. In addition to the core database, a well-designed crop breeding data management system should include a number of data analysis modules: the genealogical tree and phylogenetic analysis, molecular markers and gene discovery, data acquisition and trait dissection, selection of parents and progenies, breeding strategy, experiment design and statistical analysis, crop modeling, and dynamic gene network modeling. These data processing modules can be integrated together according to the actual needs of the crop breeding practices.

Key words: Maize breeding; Database management; Data-driven breeding technology; Data analysis; Conceptual framework

收稿日期: 2012-12-10

基金项目: 国家“863”计划“强优势玉米杂交种的创制与应用”(2011AA10A103)、国家“973”计划“玉米大豆高产优质品种分子设计和选育基础研究”(2009CB118400)

作者简介: 董春水, 博士, 旅美学者, 从事现代作物数字育种技术开发与应用研究。

E-mail: spring.water@theworldseeds.com

才卓为本文通讯作者。E-mail: caizhuo@126.com

计算机技术已经渗透到人们生活、工作的每一个角落, 数字化变成了时代的主旋律, 无论是基础科学研究还是应用技术的发展, 都在经历快速的数字化转型过程。作物育种科学承载着培育抗病虫、耐逆境、品质优并高产稳产的新品种的任务, 以满足人类社会不断增长的粮食需求, 单凭田间经验和传统技术已远远不能满足要求, 必须跟上时代步伐, 开发高效、精准数据分析和后代选择的数字化技术。

数字化是在对研究对象定性描述的基础上进一步加以定量刻画,便于进行精确评估和筛选。如观察玉米抗旱性,可根据对水分胁迫的反应以及最终产量表现,将抗性划分为高、中、低等级别,这是定性描述;在定性描述的基础上,结合分子生物学、生理学、遗传学和物候学等进一步更为精确地定量刻画,是作物数字育种技术的关键所在。定量刻画只是数字化的第一步,数字化内容涵盖了数据的分类存储和检索、数据的统计分析、各类模型的建立与应用、可视化展示和综合应用等。

早在上世纪初,生物统计学基本原理开始在作物育种中应用,这是作物数字育种技术的开端。随着计算机技术的出现,现代数字育种技术开始孕育和发展。进入 21 世纪,随着自动化采样和数据获取技术的突破,从分子水平的基因表达数据到宏观性状表现型数据都实现了质的飞跃。育种数据的存储、分析、利用成为现代数字育种技术的关键,伴随生物统计学、数据库技术、系统模拟技术和计算生物学等单学科的发展,通过将这些理论和技术有机整合,极大地推动了现代作物数字育种技术的进步。伴随着当代育种技术进步的加快,特别是生物技术的广泛应用,在作物育种中产生了信息爆炸,数据不再是单一的田间性状调查结果,还有气候、土壤等动态环境数据、分子标记等基因型数据、基因表达和代谢物动态数据、生产管理数据^[1~3]。这些数据类型复杂多变,

组成一个复杂的大数据集,为确保这些数据(如作物育种的系谱数据)不断积累和丰富,需要使用先进的数据库管理系统进行管理维护,清晰地建立各种数据之间的相互关系,便于对这些数据进行交叉分析,从中找出规律,提取有用的知识,为后代的选择和新品种的定位提供可靠的依据。同时,还要确保数据的完整性和持久一致性。

孟山都、杜邦先锋、先正达等各大种企都建有自成体系的私有数据库和管理系统,且功能十分先进与完善,存储了整个产业链条从研究部门到销售部门的各种相关数据资料,这些私有数据库系统,其结构、功能及内涵的商业机密是保密的,但可以肯定都具备海量数据的超大存储能力、复杂数据的高效分析能力、庞大系统的科学管理能力,能够为研究、管理人员提供简捷、高效、精准的服务,更好地完成相关应的育种研究,如数据的自动采集、分类、存储、分析、建模等。除这些私有数据库管理系统之外,各种商业化的数据库管理系统也在不断发展完善。

1 育种数据与管理系统

根据当代育种技术的发展现状^[4],提出一个普适性的高效育种数据管理系统框架,主要由数据库和相关的分析模块组成。图 1 为育种数据管理系统的概念模型。

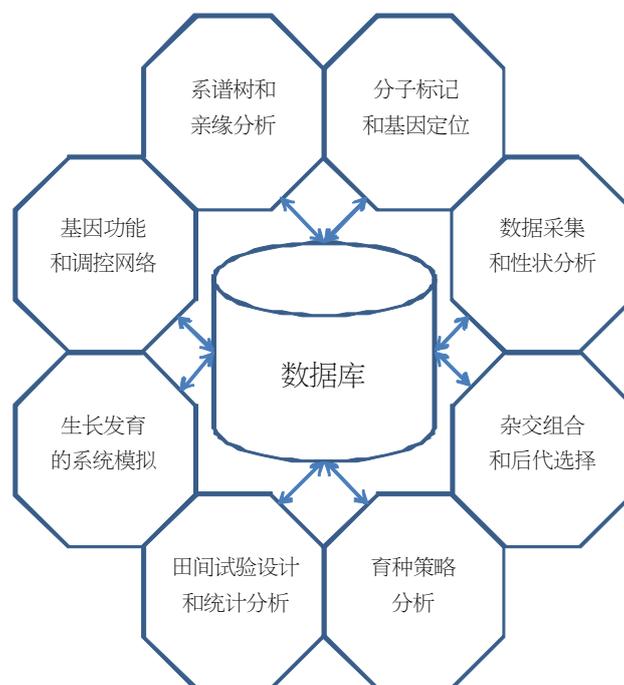


图 1 作物育种数据管理系统的概念模型

Fig.1 A conceptual framework of crop breeding data management system

作物育种数据管理系统可以包括图 1 概念模型中的一个或多个组件,数据库是系统必不可少的核心组件。除了数据库之外,各种组件都是可选的,实现特殊功能的附加组件都可随时加入系统。这是一个可扩展组合系统,除了其核心数据库之外,可以根据业务时空发展需要逐步完善。根据计算机软件开发平台的不同,可以在多个水平上实现这样的系统。一般分为单机版、客户机-服务器版和远程云计算版。

作物育种数据管理系统单机版一般个人电脑单机运行即可以满足需要,不同计算机运行的系统很少进行实时交互。适合于小育种团队,不用产生与处理大量数据,不需要强大的计算功能。

作物育种数据管理系统客户机-服务器版数据库运行于后台服务器,多台计算机通过用户界面直接与后台服务器的数据连接。系统数据库中的数据可共享,实现了用户间的实时交互。适合于作物育种团队和公司,易于实现数据的私有化。

作物育种数据管理系统云计算版是顺应当代计算机技术发展而刚刚兴起的新技术。所有数据库和应用全部运行在远程服务器上,用户通过国际互联网,使用网页浏览器就可实现随时随地对数据操作。

目前,玉米育种可直接应用的主要商用育种数据管理系统有 PRISM(植物研究信息共享管理系统 <http://www.teamcssi.com/index.html>)、AGROBASE Generation II (<http://www.agronomix.com/>)、农博士育种家 (<http://www.nbs.net.cn/yzjaa.htm>)、和 Phenom-Networks (<http://www.phenome-networks.com/>)。PRISM是 Central Software Solutions 公司旗舰产品,被多国大型种企应用于支持玉米、大豆、棉花、水稻等多种作物;PRISM使用中心共享数据库,数据库可包含上百万条记录,支持 600 多用户同时使用,便于研究人员共享试验数据;客户端运行在 Windows 操作系统下,中心数据库则运行 SQL 服务器上,单用户版本使用 SQL Express 数据库软件。AGROBASE Generation II 是 Agronomix Software Inc 公司产品,有单机版和客户机-服务器版;1990 年发布以来,被 40 多个国家农学和育种家采用,适于玉米、大豆、高粱等作物育种和品种测试,在 Windows 系统上运行;客户机-服务器版使用 MS SQL SERVER 作为数据库平台,核心系统提供数据管理、试验设计和数据分析,根据需要可以添加各种系统模块。农博士育种家是北京中农博思公司产

品,包括育种数据采集、管理、分析等功能,适用于玉米、小麦、水稻等多种作物;使用 SQL SERVER 数据库服务器管理数据,支持海量数据的存储。Phenom-Networks 以色列希伯来大学创建的生物信息学公司(Phenom-Networks)产品,是世界上第一个使用云计算的育种软件。目前,Phenom-Networks 通过网页提供简单的育种数据管理、表型数据的统计分析和 QTL 定位等遗传分析。

2 数据类型与数据库

2.1 数据类型

作物育种过程中具有不断产生数据的特点,可以根据对数据的实际需要对其进行多种形式的定义分类。在实际应用上,根据数据产生的对象、数据的时间特性、数据的来源和生物学尺度进行分类,分类的科学实用与否会对存储、分析和利用产生大的影响。

2.1.1 产生对象

根据数据产生的对象可以将数据分为表现型、基因型、环境影响、田间管理等,表现型数据来自对植物本身的观测;基因型数据是来自遗传物质 DNA,可以是分子标记数据也可以是物理测序数据;环境影响数据包括气候资料和土壤数据及病虫害危害数据;田间管理数据包括播种、灌溉、施肥、收获等相关数据。

2.1.2 时间特性

根据数据产生的时间特性可以将数据分为动态时间序列和静态非时间序列。叶面积为群体水平上的动态时间序列;平均产量为群体水平上的非时间序列。

2.1.3 数据来源

根据数据的来源可以将数据分为原生数据、次生数据和预测数据,原生数据直接来自田间和试验的观测;次生数据是通过统计分析或其他复杂的计算形成;预测数据是通过模型运算产生。

2.1.4 生物学尺度

根据观测数据的生物学尺度不同,将数据分为宏观和微观,是一连续变化的系统^[9],一般划分为基因、DNA/RNA/蛋白质/代谢物、细胞器、细胞、组织、器官、个体、群体和群落等。从分子水平的基因一直到宏观水平的生态群落,都有可能对作物育种的某个环节出现并对育种决策起作用。

2.2 数据之间的相互关系

现代数据库技术是高效管理这些数据的必备

工具,需要理清各种类型数据的特点及其相互关系,创建高效的关系数据库。不同领域的科学家提出了各种不同数据模型,用来存储这些相关的数据,作为育种数据管理系统的核心模块,为各种数据的交叉分析提供了可能^{6,7}。图2提供了一个概念性的数据关系结构,包括基因型、表现型、育种信息、性状信息和环境信息5大类数据。其中,每个方框都代表一组具有特殊属性的数据,可以存储在一张或多张

数据表中,以便于与其他数据协同使用,为育种决策提供支持。箭头标示了关联信息的影响关系,如环境信息通过基因型产生一组表现型,育种信息和性状信息都可以直接施加于基因型,从而产生新的基因型。这些数据之间都是相互影响和关联的,数据之间的关系并非一成不变,这些数据的总和构成了作物育种数据管理系统的核心数据库。根据每个育种公司和育种计划,这些关系需要重新调整,便于适

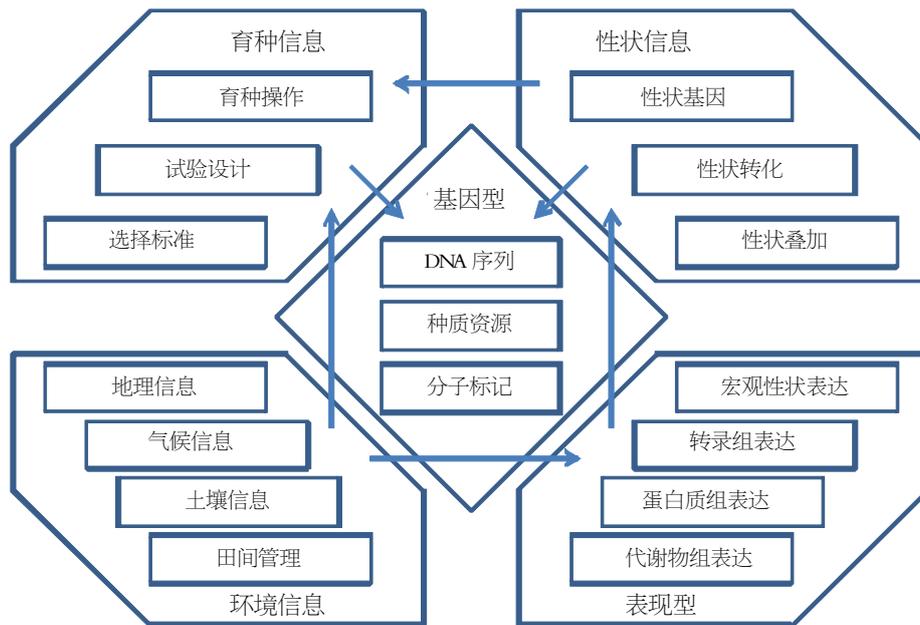


图2 作物育种各种数据的相互关系

Fig.2 Possible relationships between different kinds of data in crop breeding

应特殊需要。

基因型是作物育种的基石,录入了自交系及种质资源的详细信息、分子标记和DNA测序,未来还会包括DNA修饰的表观遗传信息等。这些信息描述了种质资源的根本特性,用于研究系谱结构和材料之间的亲缘关系,为亲本选择提供必不可少的参考。对应表现型数据可以用于数量性状的基因定位、发现新基因和建立基因调控网络,为转基因性状育种服务。

表现型是基因型在不同环境条件下的具体表现,录入了包括宏观性状如产量、株高、叶面积的动态变化,也包括基因表达的mRNA、蛋白质和代谢物等微观性状的动态变化信息。这些数据随时间、环境和基因型而变化,所以一个基因型可有許多表现型数据集与之对应。

育种信息包括了杂交组配、亲本和后代选择以及田间试验设计的信息,录入了杂交组配、后代选择的规则和评估材料的试验处理信息,来源于育种家

经验和知识的积累凝练,再为育种家进行数据分析和材料选择提供支持。

性状信息是多数公共育种数据管理系统所没有考虑的数据,录入了包括各种各样的基因信息,如QTL定位和转基因事件以及不同性状或基因之间的叠加信息,是转基因育种所必备的信息。

环境信息是地理位置与育种相关资料的属性,录入了地理坐标、气候、土壤、耕作、管理等信息,综合起来构成一个育种试验环境,基因型在这个环境中的表现构成一个表现型数据集。由于生物体的复杂性和对环境反应的差异性,决定了基因型在不同环境下的表现不尽相同,这就是基因型与环境的互动。

2.3 玉米育种相关的公共数据库

美国有多个比较完善的玉米基因序列和遗传资讯数据库,这些数据库对育种家研究不同性状的遗传机制和改良玉米自交系有极大帮助⁸。

MaizeGDB(<http://www.maizegdb.org/>)提供了包括

遗传图谱、基因序列、QTL、等位基因等丰富的遗传和基因组信息,基因详细信息可直接在线检索,检索结果包含许多有用的交叉链接,相关文献十分丰富。

MaizeSequence (<http://www.maizesequence.org>)存储了玉米基因组测序项目(Maize Genome Sequencing Project)完成的最新基因组序列,可在线直接查询 B73 自交系的基因序列。

Gramene (<http://www.gramene.org>)遗传多样性数据库存储了水稻、玉米、小麦等禾谷类作物的基因组信息。由于禾本科较近的亲缘关系,基因组之间存在很多相似之处,研究一个作物的基因、调控网络和表现型,可以直接或间接地应用到其他近缘作物。

Panzea (<http://www.panzea.org/index.html>)汇集了多个美国自然科学基金玉米项目(<http://www.maize-genetics.net>)产生的基因型、表现型和分子标记数据资讯,如 NAM 群体(Nested Associated Mapping Population)的基因型和表现型数据。除了提供公共检索服务外,还允许下载完整的数据库。

3 数据采集与性状分解

在田间如何及时、快速、高效地获取玉米生长发育的准确数据一直是作物育种的瓶颈。近年来,育种数据采集方法有了长足的发展。基因分型方法的进步使分子标记甚至 DNA 序列的测定变得实用简捷,特别是美国 Illumina 公司(<http://www.illumina.com>)研制出多种高通量的基因分型设备和 DNA 测序设备,使分子标记的测定过程步入自动化,为基因型数据的快速积累起了关键作用,基因型数据的迅速积累呈指数增长,远远超过了其他类型的数据积累;随着以遥感、图像处理等为代表的表现型数据的测定获取技术以及以自动化气象站等为代表的自动化监测设备的完善^[9,10],使表现型数据、环境数据的积累也迅速跟上基因型数据的同步积累,形成了海量信息。然而,育种实践中往往要对各项数据进行综合分析才能得出全面、准确的结论。信息爆炸致使这些分析变得更加复杂、困难。

玉米生长发育和产量形成是一个复杂过程,育种目标性状也必然是多个因素的综合作用结果。因此,需要将复杂性状分解成简单性状,通过分子育种技术等新育种技术手段研究这些简单性状的遗传结构,便于实现多性状的综合育种目标。性状分解不是随意的,是随现有技术水平和人们对性状的认识水平而定的,如玉米产量是一个高度综合的性状,产量性状的分解可以是传统意义上的产量构成因子(穗

粒数、穗数和粒重),随现代技术发展,又可把产量细化分解为多个新因子性状。杜邦先锋公司研发的玉米穗成像专利技术将产量性状分解为单个体粒性状、单个果穗性状和群体性状(<http://www.faq.s.org/patents/app/20090046890>);德国 LemnaTec 公司(<http://www.lemnatec.com>)开发出高通量数据采集系统,使用图像处理技术来快速获取必要的作物生长发育数据,如叶面积、叶绿素含量、茎粗、株高、生长速率、节间长度、生物量、叶片卷曲等。

这些专用模块可以组合到作物育种数据管理系统中作为数据获取的端口,为育种家输入、获取详尽的相关数据,便于为大量筛选转基因材料、研究基因功能、获取表现型数据和常规育种后代选择提供支持。

4 田间试验设计与统计分析

近年来,育种试验设计长足发展,为育种遗传材料的评估提供了有效的统计学保证,试验数据的可靠性和可用性大幅度提高。玉米早代育种材料数量庞大,由于植株高大、土地资源所限,测试只能在少数地点进行,试验设计多采用高度不平衡设计,除对照品种,测试材料基本没有重复。常用的试验设计方法有格子方设计(lattice design)、 α -设计(α -design)、增强设计(augmented design)、p- rep 设计(p- rep design)、增强 p- rep 设计(Augmented p- rep design)^[11,12]。

做好试验设计并采集到高质量的数据只是成功的第一步,对采集的数据分析、比较和选择育种材料才是关键。在育种上最常用的数据分析模型是混合模型^[13],在这些模型中,育种材料被当作随机变量,对育种材料的估计采用的是最佳线性无偏预测(Best linear unbiased prediction, BLUP)。对试验中的多个对照材料的处理也十分重要,处理合理可以提高试验分析的比较精度^[14]。常用统计软件有 R 语言、AS-REML、SAS、S-Plus、Minitab、MATLAB 等软件工具。另外,各种多元统计分析方法如主成分分析、聚类分析、判别分析、决策树等也十分重要。近年,人工智能的算法也被应用到育种数据分析,如人工神经网络、自组织图、遗传算法等^[15]。

5 分子标记与基因定位

分子育种已经成为作物育种中必不可少的一种现代技术,它基于作物基因组的分子标记的基因定位和后代选择技术。在实际应用时,可以是对一个或多个育种群体通过大量的田间测定,获得在不同生

长环境下的表现型数据集,同时进行基因型测定,获得尽可能详尽的分子标记数据;然后通过各种 QIL 定位方法,在性状表现型和分子标记之间建立起关联。一个受多位点控制的数量性状与基因组中多个区域的分子标记产生紧密的关联关系,这只是基因定位的一种途径,还可以通过关联分析和全基因组关联分析或最近提出的 QIL-seq 方法^[16]建立起表现型和分子标记或基因组区域的联系。通过精细基因定位方法,缩小搜索范围,找到起关键作用的基因。

QIL Cartographer^[17] 对自交系的杂交后代群体 QIL 定位,并包含强大的图形功能,可以将结果用图表形式显示出来,也可以使用多种格式输入和输出数据。它包含有单分子标记分析法、区间定位法、合成区间定位法、贝叶斯区间定位法、多区间定位法、多性状定位法、分类数据的定位方法等数据分析方法。

Biomercator^[18]用于对多个 QIL 分析结果进行综合的 Meta 分析,帮助育种家对数量性状的多个 QIL 研究结果进行综合分析,便于找到的对应基因。

TASSEL^[19]用于评估性状关联、进化格式、连锁不平衡关系的软件,可为关联基因定位提供新的高效的统计方法,如一般线性模型(GLM)和混合线性模型(MLM)^[20];能够处理广泛的插入删除(indels)标记。许多软件不考虑这种类型分子标记的遗传多样性,但是这种标记在一些物种十分普遍。但是,现在还没有十分成熟的基因组选择软件用于作物育种,CIMMYT 开发出使用高密度分子标记和系谱数据进行基因组选择的 R 软件包(http://genomics.cimmyt.org/BLR_1.1.tar.gz) 实现了不同的基因组选择方法。Wimmer 等^[21] 开发出 synbreed (<http://cran.r-project.org/web/packages/synbreed>)R 软件包,用于多种基因组模型的数据拟合和交叉验证等。

在实际育种数据管理系统中,这些方法需要有机地整合,才能使各种数据可以无缝地与数据分析模块的这些算法融合,从而进行 QIL 或基因定位,真正实践为育种服务。

6 系谱树与亲缘分析

自交系系谱承载着其亲缘关系信息,对后续育种的亲本选择具有非常重要的参考价值。美国各大育种公司和国际研究所都有保存完好的系谱记录,可以轻易绘制出一张详细的系谱树,供育种家研究和参考。中国缺乏系统、完整的自交系血缘系谱信息,就更需要使用现代分子标记技术追溯自交系间

亲缘关系^[22,23],为后续的亲本选择及组合选配提供参考。伴随国内商业育种体系逐步形成与完善,种业公司的育种团队可以详细的记录自交系的亲本信息逐步积累,也会为今后育种技术的发展奠定基础。

2000 年以来,Structure(<http://pritch.bsd.uchicago.edu/structure.html>)软件历经逐年改进,形成了完善的软件结构和功能^[24],使用多位点基因型数据研究群体结构,可以用来推断是否有异质种群存在、判定个体所属种群、研究种群混杂区域、辨识迁徙和混杂的个体、估计群体中等位基因的频率等,还可以使用各种分子标记数据,如 SNP、RFLP、AFLP 和微卫星标记等。

7 杂交组合与后代选择

杂交亲本的选择是能否成功培育出优良自交系的关键。在现代玉米育种中,常规做法是选择一个符合育种目标的优秀自交系作为改良的遗传基础,再选择一个或多个来自同一个杂种优势群内可以克服其缺点的其他自交系来加以改良,避免因亲本血缘关系混杂降低杂种优势的潜力。由于杂种优势群复杂,杂优模式多样,又有很多自交系含有墨西哥、CIMMYT 热带种质,为亲本选择增添很多困难,育种家尤其要注意通过准确的判别材料亲缘关系选择亲本。用于杂交亲本和后代选择的成熟商业软件不多,还主要靠育种家根据各个性状的表现权衡利弊进行选择。可以采用聚类分析、主成分分析、判别分析、决策树、自组织图、人工神经网络等多种多元分析或人工智能方法,辨别出优良的后代,从而使后代的选择更加规范和有据可循,实现决策过程的自动化,而非育种家凭自己的一己判断而定^[25]。利用全基因组选择法,可以在后代播种到田间之前提前对可能的表现进行预测,舍弃预测值很低的家系后代,可以大大提高选择的定向性和目标性。

8 生长发育系统模拟

作物品种在多变的环境中的生长发育遵循一定的物理学和生理学基本规律,其生长发育过程可以使用数学方程进行量化,因此,可以组建作物生长发育模拟模型,实现作物在不同环境条件下生长发育和产量形成的模拟。自从上世纪 70 年代开始,在荷兰和美国发展起来的作物模拟技术经历了漫长的发展,主要用于农场的生产管理。

近年来,作物模拟技术有了长足发展,融入作物遗传信息,使模型的预测精度得以提高,人们逐步开

始在作物育种中使用作物模拟模型,研究基因型与环境的相互作用(G x E interaction);利用玉米生长发育模型对生长环境进行分类,进一步实现玉米杂交种定位,为产品的销售服务^[27];利用作物模型研究得出美国玉米带玉米产量逐年增加的内在原因是垂直下扎较深的根系可能是高密度种植条件下增产的内在原因^[28]。目前广发应用的玉米生长发育模型有 CERES-maize 模型和 APSIM-maize 模型。

CERES-maize 模型^[29,30](<http://dssat.net/>)多作为研究工具,输入参数包括每日天气数据,如最高和最低气温、降雨量和太阳辐射;分层的土壤水分特征参数,如土壤饱和含水量、田间持水量、萎蔫点土壤含水量、土壤容重等;田间管理措施,如播期、播深、灌溉时间和灌溉量、施肥时间和用量等。作物遗传参数用来描述玉米植株的生长发育和产量形成各个生理过程与各种环境变量之间的关系,能够模拟并输出土壤的水分动态、植株的生育期动态、干物质的生成和分配动态、产量和产量因子的动态变化等。

APSIM-maize 模型^[31](<http://www.apsim.info/Wiki/>)是在 CERES-maize 模型的基础上发展来的,在生理过程模拟方面进行了改进提高,其基本结构和功能相似。

目前,如何在作物科学中有效运用作物模型是科学家仍在探索的问题,其中最主要原因是使用作物模型工具可以完成常规手段难以完成的研究工作。如研究玉米在过去 100 年气候条件下的表现、研究对不同水分胁迫环境的反应、研究对各种土壤类型的反应等,为新品种的选育提供高效的辅助手段。这些研究即使常规手段仍然可行,但人力、物力和费用也将十分惊人。

9 育种策略分析

育种策略决定着育种计划的成败,包括但不限于种质资源的选择、育种方法、选择环境以及何种数据需要采集等^[32]。用于育种策略研究的模型有 QU-GENE 育种模拟系统(<http://www.uq.edu.au/lcfs/qugene/>),由澳大利亚昆士兰大学开发^[33],研究在多轮选择和杂交的育种过程中遗传系统的特性,包括两层结构:基因型-环境系统引擎和应用模块。用户只需在可视界面完成 3 个步骤,详细定义遗传-环境系统,定义杂交和选择计划,运行模拟即可直观地显示相关结果。QU-GENE 和玉米作物模型结合用于研究抗旱育种,首先通过改变作物模型中不同的遗传参数(性状),模拟优良品系在多年多点的不同

同生长发育环境下的产量表现;然后,遗传参数和产量在不同环境条件下的等产线图,用 QU-GENE 来确定不同遗传参数对育种目标的重要性,从而提出育种策略并应用到玉米抗旱育种的实践中^[34]。

10 基因功能与调控网络

任何基因在生物体中起作用都不是孤立而是参与一个或多个基因调控网络进而实现其功能。因此,基因功能和基因调控网络的研究在转基因育种中的作用不言而喻。通过用数学方程式表述这些基因和基因网络,开发基因网络动态模拟模型,进而模拟基因表达的动态变化,来预测宏观性状表现型的状态。过去相关研究主要应用在微生物和动物,直到近年才逐步应用到植物基因及其功能研究。Welch 等^[35]率先研究拟南芥开花期基因网络。Dong 等^[36]将这种方法应用到玉米开花期网络研究,创建了一个包含 4 个重要花期基因的动态网络模型来预测玉米开花期。

由于作物生理过程复杂繁多,针对目标性状的基因网络动态模型需要具有可扩展性,利用存储于核心数据库的数据,当模型扩展后可以方便地调用以前的数据,对扩展后的模型进行验证和再训练,确保在模型扩展时不牺牲模型的兼容性,还需要将必要的构件如动态模型本身、动态模型的训练或优化以及模型的输出有机组合起来,组装到育种数据管理平台中。

参考文献:

- [1] Tinker N A, Yan W. Information systems for crop performance data[J]. *Can. J. Plant Sci.*, 2006, 86: 647-662.
- [2] Eathington S R, Crosbie T M, Edwards M D, et al. Molecular Markers in a Commercial Breeding Program[J]. *Crop Science*, 2007, 47: S-154-S-163.
- [3] Heffner E L, Sorrells M E, Jannink J L. Genomic selection for crop improvement[J]. *Crop Sci.*, 2009, 49: 1-12.
- [4] Dong C-S, Cai Z. Current status and perspectives of maize breeding technologies[J]. *Journal of Maize Sciences*, 2012, 20:1-9.
- [5] Mochida K, Shinozaki K. Advances in omics and bioinformatics tools for systems analyses of plant functions[J]. *Plant Cell Physiol.*, 2011, 52: 2017-2038.
- [6] Du C G, Buckler E, Muse S. Development of a Maize Molecular Evolutionary Genomic Database[J]. *Comparative and Functional Genomics*, 2003, 4: 246-249.
- [7] Heckenberger M, Maurer H P, Melchinger A E, Frisch M. The Plabsoft database: a comprehensive database management system for integrating phenotypic and genomic data in academic and commercial plant breeding programs[J]. *Euphytica*, 2008, 161: 173-179.
- [8] Sen T Z, Harper L C, Schaeffer M L et al. Choosing a genome browser

- for a Model Organism Database: surveying the Maize community[J]. Database, 2010: baq007, doi: 10.1093/database/baq007.
- [9] Vina A, Gitelson A A, Rundquist D C, et al. Monitoring maize phenology with remote sensing[J]. *Agronomy Journal*, 2004, 96: 1139–1147.
- [10] Elwadie M E, Pierce F J, Q J. Remote Sensing of Canopy Dynamics and Biophysical Variables Estimation of Corn in Michigan[J]. *Agron. J.*, 2005, 97: 99–105.
- [11] Smith A B, Lim P, Cullis B R. The design and analysis of multi-phase plant breeding experiments[J]. *Journal of Agricultural Science*, 2006, 144: 393–409.
- [12] Williams E R, Piepho H P, Whitaker D. Augmented p-rep designs[J]. *Biometrical Journal*, 2011, 53: 19–27.
- [13] Smith A B, Cullis B R, Thompson R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches[J]. *Journal of Agricultural Science*, 2005, 143: 449–462.
- [14] Müller B U, Schutzenmeister A, Piepho H P. Arrangement of check plots in augmented block designs when spatial analysis is used[J]. *Plant Breeding*, 2010, 129: 581–589.
- [15] Omella L, Cervigni G, Tapia E. Application of machine learning in breeding for stress tolerance in maize[M]. In B Venkateswarlu et al. (Eds) *Crop stress and its management: perspectives and strategies*, DOI 10.1007/978-94-007-2220-0_5.
- [16] Takagi H, Abe A, Yoshida K et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from twd bulked populations[J]. *The Plant Journal*, doi:10.1111/tpj.12105.
- [17] Wang S, Basten C J, Zeng Z-B. Windows QTL Cartographer 2.5[M]. Department of Statistics, North Carolina State University, Raleigh, NC. 2012(<http://statgen.ncsu.edu/qtlcart/WQTLCart.htm>).
- [18] Sosnowski O, Charcosset A, Joets J. BioMercator V3: an upgrade of genetic map compilation and quantitative trait loci meta-analysis algorithms[J]. *Bioinformatics*, 2012, 28: 2082–2083.
- [19] Bradbury P J, Zhang Z, Kroon D E et al. TASSEL: Software for association mapping of complex traits in diverse samples[J]. *Bioinformatics*, 2007, 23: 2633–2635.
- [20] Yu J, Pressoir G, Briggs W H et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness[J]. *Nature Genetics*, 2006, 38: 203–208.
- [21] Wimmer V, Albrecht T, Attinger H-J et al. synbreed: a framework for the analysis of genomic prediction data using R[J]. *Bioinformatics*, 2012, 28: 2086–2087.
- [22] Semagn K, Magorokosho C, Vivek B S et al. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single nucleotide polymorphic markers[J]. *BMC Genomics*, 2012, 13: 113.
- [23] Wen W-W, Araus J L, Shah T et al. Molecular characterization of a diverse maize inbred line collection and its potential utilization for stress tolerance improvement[J]. *Crop Sci.*, 2011, 51: 2569–2581.
- [24] Pritchard J K, Stephens M, Donnelly P J. Inference of population structure using multilocus genotype data[J]. *Genetics*, 2000, 155: 945–959.
- [25] Hubisz M J, Falush D, Stephens M et al. Inferring weak population structure with the assistance of sample group information[J]. *Molecular Ecology Resources*, 2009, 9: 1322–1332.
- [26] Azinzadeh M, Aniri R, Davoodi-Bojd E, et al. Computer aided selection in breeding programs using genetic algorithm in MATLAB program[J]. *Spanish Journal of Agricultural Research*, 2010 8(3): 672–678.
- [27] Loeffler C M, Wei J, Fast T, et al. 2005. Classification of maize environments using crop simulation and geographic information systems. *Crop Sci.*, 2005, 45: 1708–1716.
- [28] Hammer G L, Dong Z-S, McLean G, et al. Can changes in canopy and/or root system architecture explain historical maize yield trends in the U.S. Corn Belt[J]. *Crop Sci.*, 2009, 49: 299–312.
- [29] Jones C A, Kiniry J R. CERES-Maize: A simulation model of maize growth and development[M]. Texas A&M University Press, College Station, Texas, 1986.
- [30] Jones J W, Hoogenboom G, Porter C H, et al. DSSAT cropping system model[J]. *European Journal of Agronomy*, 2003, 18: 235–265.
- [31] Keating B A, Carberry P S, Hammer G L, et al. An overview of APSIM, a model designed for farming systems simulation[J]. *European Journal of Agronomy*, 2003, 18: 267–288.
- [32] Bänziger M, Edmeades G O, Beck D, et al. Breeding for drought and nitrogen stress tolerance in maize: from theory to practice[M]. Mexico, D.F.: CIMMYT, 2000.
- [33] Podlich D W, Cooper M Q. GENE: A platform for quantitative analysis of genetic models[J]. *Bioinformatics*, 1998, 14: 632–653.
- [34] Messina C D, Podlich D W, Dong Z-S, et al. Yield-trait performance landscapes: from theory to application in breeding maize for drought tolerance[J]. *Journal of Experimental Botany*, 2011, 62: 855–868.
- [35] Welch S M, Dong Z-S, Roe J L, et al. Flowering time control: gene network modelling and the link to quantitative genetics[J]. *Australia Journal of Agricultural Research*, 2005, 56: 919–936.
- [36] Dong Z-S, Danilevskaya O, Abadie T, et al. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling[J]. *PLoS ONE*, 2012, 7(8): e43450.

(责任编辑:李万良)