浙江大学学报(农业与生命科学版) 42(1):00～00,2016
Journal of Zhejiang University (Agric. & Life Sci.)
http://www.journals.zju.edu.cn/agr
E-mail:zdxbnsb@zju.edu.cn
DOI:10.3785/j.issn.1008-9209.2015.11.102

作物育种相关数据及大数据技术育种利用

樊龙江[1*], 王卫娣[1], 王斌[2], 叶楚玉[1], 舒庆尧[1], 张辉[3]
(1.浙江大学作物科学研究所/生物信息学研究所/IBM 生物计算实验室,杭州 310058;2.河南科技学院生物工程系,河南 新乡 453003;
3.中华人民共和国科学技术部中国农村技术开发中心,北京 100045)

摘要 从18世纪首次获得人工杂交种到如今基因工程育种,作物育种技术发展迅速,同时几百年的育种历程积攒了大量育种数据,特别是近年来伴随高通量测序技术的发展,产生了海量作物育种相关基因及其表达数据,形成了育种大数据.2012年以来在商业、信息技术等领域发展迅猛的大数据技术,致力于解决大数据采集、存储及处理等壁垒,并在其他领域的应用初露睥睨.本文利用创新方法TRIZ(theory of inventive problem solving)流分析技术,综合分析了育种领域已有资源和目标达成的矛盾问题,提出大数据育种技术应用于作物育种的创新方案,明确了将大数据技术应用于育种领域的框架和实现目标.提出了基于大数据理念的育种技术,拟采集和整合已有育种数据资源,实现数据自动采集等,从而能够平衡育种数据膨胀/利用和育种需求产生的矛盾;构建基于大数据技术的育种数据信息化平台,实现作物育种方法理念的创新,为广大育种工作者提供数据支撑和一个育种新途径;为解析生物学数据与目标农艺性状的关系提供信息,加快育种现代化的进程.

关键词 作物育种; 大数据技术; 育种方法; 大数据育种技术; 创新方法
中图分类号 S5-3 **文献标志码** A

Crop breeding-related data and application of big data technologies in crop breeding. Journal of Zhejiang University (Agric. & Life Sci.), 2016,42(1):000－000

Fan Longjiang[1*], Wang Weidi[1], Wang Bin[2], Ye Chuyu[1], Shu Qingyao[1], Zhang Hui[3] (*1. Institute of Crop Science & Institute of Bioinformatics & IBM Biocomputational Laboratory, Hangzhou 310058, China; 2. Department of Bioenegineering, Henan Institute of Science and Technology, Xinxiang 450033, Henan, China; 3. China Rural Technology Development Center, Ministry of Science and Technology of the People's Republic of China, Beijing 100045, China*)

Summary Since the first artificial hybrid was created in 1719, significant developments such as transgenic approach have been seen in the methods for crop breeding in recent hundreds of years. A lot of breeding-related data have been collected up to now. The big data technology was developed recently and has been successfully used in economics, IT (information technology) and other fields. With the increasing expansion of data in crops breeding, it becomes extremely necessary for breeders to take advantage of existing data in terms of efficient breeding technology, especially for the information generated from next-generation sequencing which could newly reach

基金项目:中华人民共和国科学技术部创新方法工作专项资助项目(2015IM010400);中华人民共和国科学技术部科技基础性工作专项资助项目(2013IM030700).

* **通信作者**(Corresponding author):樊龙江(http://orcid.org/0000-0002-2892-7102), Tel:＋86－571－88982730; E-mail:fanlj@zju.edu.cn

收稿日期(Received):2015－11－10;**接受日期**(Accepted):2015－11－25;**网络出版日期**(Published online):2015－12－15
URL:

terabytes of data in a sequencing platform in one hour. In this study, we proposed a conceptual framework for big data-based crop breeding approach after we analyzed genetic information flow of crop breeding program using an innovation tool, TRIZ (theory of inventive problem solving). The expected new breeding technique tends to collect all breeding-related data (including data from phenotype, environments, references to molecular markers and sequences) for target crops and set up an automatic approach to collect breeding-related trait data. The technique will include a computer system to analyze the data and a human machine interface for users (breeders). We believe that big data has the potential to revolutionize the breeding of crops and the big data-based breeding approach is our future of crop breeding programs.

Key words crop breeding; big data technology; breeding method; big data-based breeding approach; innovation method

在人类早期简单的种植和采收活动中,就开始孕育作物驯化育种的思维.中国在周朝已形成不同播期和熟期的作物品种概念(见《诗经》"黍稷重穋""稙穉菽麦").在源于西欧的近代育种技术和理论出现之前,作物育种都是通过天然杂交和变异产生一些符合人类生产需求的作物品种.1719 年托马斯·费尔柴尔德(Thomas Fairchild)以石竹科植物为材料首次获得人工杂交种,随后奈特和库尔特分别于 1823 年和 1843 年用豌豆和谷禾类作物进行人工杂交育种.而自孟德尔定律在 20 世纪初提出后,遗传学、分子生物学、生物统计学等学科的建设和快速发展,使得人工作物育种开启了新篇章.自新一代测序技术高速发展以来,作物育种研究产生了海量多种类型的数据,整合和最大化利用这些生物学数据,无疑对现代育种研究具有不可估量的重要意义.因此,构建作物育种相关数据库利用平台,致力创造大数据背景下的育种技术,可以平衡育种数据膨胀和育种需求产生的矛盾,从而实现育种数据数字化平台建设,也为广大育种工作者提供数据支撑,同时也为广大生物学家了解生物学数据与目标性状的关系提供渠道.

1 育种相关数据及其数量估计

农作物育种是一项复杂的系统工程,涉及种质资源鉴定与创新、新基因发掘、育种技术、品种培育、种子生产及其产业化等.世界主要国家均把农作物育种及其产业作为提高国家竞争力的重要战略选择,竞相投入大量的人力、物力和财力研发现代育种技术,培育新品种,抢占制高点,促进作物育种及种业的持续发展.农作物品种选育呈多元化发展态势.高产是新品种选育的永恒主题;品质改良是新品种选育的重点;病虫害抗性是新品种选育的重要选择;非生物逆境是新品种选育的重要方向;养分高效利用是新品种选育的重要目标;适宜机械化作业是新品种选育的重要特征.

现代科学技术持续创新引领农作物育种发生深刻变革.新技术的应用,包括生物组学、生物技术、信息技术、制造技术等现代科学技术飞速发展,不断渗入农作物育种各个层面,催生了新型的农作物育种体系.例如,表型组学和基因组学技术不断深化种质资源鉴定与评价,如采用先进的移动式激光 3D 植物表型成像系统,高通量测序技术;新基因挖掘与基础研究取得明显进展;前沿技术引领育种方向,育种科技创新呈高新化,以转基因、分子标记、单倍体育种、分子设计、基因组编辑技术、全基因组选择技术等;现代信息与智能化技术广泛用于农作物育种.围绕新品种选育的实际过程,以性状数据采集和处理分析为核心,以育种过程管理为基础,实现对育种的信息化管理和数据的科学化分析,全面提高育种的管理水平和数据处理能力.

现当代育种技术(尤其是生物技术的应用)的发展,使得作物育种数据呈现了信息爆炸,所获得的育种数据不局限于单一的田间性状调查结果,同时还存在土壤、气候、水分等动态环境,影响数据、基因表达、分子标记等基因型数据和代谢物动态数据、以及生产管理数据[1].而数字化育种,滕海涛等[2]将其定义为"通过对广泛的动态育种数据的标准化管理和分析,对育种材料综合属性进行自动数据处理,对育种材料进行遗传距离和类群分析、杂种优势预先判定,对育种有关的环境因素、田间试验等数据加以考虑,按需选择育种结果".由于育种数据的膨胀,借鉴这种育种方式理念,提高育种的目标性、准确性和育种效率,育种过程中大数据管理和利用呼之欲出.

国外很多跨国种业公司已然意识到育种数据不可估量的价值,并且已经加以利用.例如,董春水

等[1]提到,“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,能够为研究和管理人员提供简捷、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”.

1.1　育种相关数据

1.1.1　基因组测序数据

1977 年,Sanger 等发明了“末端终止法 DNA 测序”的应用,使大规模、自动化 DNA 测序得以发展.1988 年到 2001 年,焦磷酸测序技术从发展到成熟,推进低成本的 DNA 测序技术发展.2005 年,高通量测序技术开始萌发.如今,DNA 测序技术已然是成熟的低成本、高效率、高质量的生物研究技术[3].

基因组序列对生物学家用于揭示物种生命本质和利用生物资源有着重要意义,而关于模式植物以及一些重要农作物的基因组信息的披露,可以促进育种进程向前推进.随着测序技术的不断推进,目前已完成了部分农作物的基因组测序,例如水稻、高粱、玉米、大豆、大麦、小麦、棉花、小米、马铃薯等.

1.1.2　转录组测序与分子标记数据

转录组广义上指的是某一特定生理条件下,细胞内所有转录产物的集合,包括 mRNA、rRNA、tRNA 及非编码 RNA(non-coding RNA);狭义上指的是所有 mRNA 的一个集合.转录组是研究基因表达的一个主要手段,因为转录组是可以连接起基因组上的遗传信息与具备生物功能的蛋白质组的一条必然纽带,基于转录水平的调控则是目前研究领域涉及最多的,也正是生物体中最重要的一种调控方式.

转录组测序的研究对象为特定细胞在特定生理功能状态下所有可能转录出来的 RNA 总和,主要包括 mRNA 和非编码 RNA .转录组的研究是基因功能及其结构的重要基础和出发点,随着新一代高通量测序技术的发展,已经使高通量转录组测序,即 RNA-seq(RNA sequencing),实现全面快速地获得特定物种、特定组织或特定器官在特定状态下产生的几乎所有转录本序列的信息.转录组测序及序列利用这一重要研究手段已广泛应用于基础科学研究、临床诊断和各种药物研发等领域.

DNA 分子标记作为在作物种质资源中发现遗传差异的新方法.近年来,育种家们已经在各种作物中发现了许多的分子标记,利用这些标记实现了高密度永久分子遗传物理图谱的构建,从而为标记辅助选育新基因或 QTL 识别提供依据.

1.1.3　作物表型检测数据

自近现代人工育种兴起以来,几乎所有的育种项目,都会在实验室、温室或者大田进行表型检测.最开始依靠人为观察记录,而随着光影科技和数字化发展,特别是显微镜的进步,现在大部分表型数据已经实现了自动化图像记录[4].植物表型数据标准化和数据储存处理对于基因组表型关联分析有着重要意义[5].近百年来记载的作物表型数据,包括文字记载和采集的图形文件,对作物育种研究有着重要意义.

孟山都和杜邦先锋公司目前都已采用温室自动化技术,包括对温室中的每一株植物进行编号,定期使用传送带将植物送到数据采集室,通过多方位和多光谱自动照相技术采集相关数据,再通过图像分析技术建立相应的生长模型,以还原植物在特定处理条件下的生长情况.

1.1.4　田间数据与农业环境数据

田间数据的采集从人工采集到现在依靠移动设备和微型无人机,育种数据的客观性也在逐渐提升.依靠便携式测量仪器、红外传感器和无人机摄影技术,目前已经实现了对大田作物生长数据全方位的检测.其中包括作物水分含量,生长时期,田间光照,温度,湿度,土壤水分等,甚至可时刻预测作物病虫害感染状况.

农业环境数据的检测,对作物品种推广有着重要作用,由于作物生长环境的复杂多变,不同地域光照、温度、土壤水分含量、空气组成成分、病虫害等条件存在差异,同一地区不同地块也存在土壤肥力和小气候的不同.为了实现种质资源的地域匹配,2013 年,孟山都公司耗近 10 亿美元收购了气候公司(Climate Incorporation),2014 年又收购了 Solum 公司的土壤分析板块,以增强其在土壤气候观测和模拟以及农业数据模型方面的实力.孟山都还花费 2.5 亿美元收购了精确播种公司(Precision Planting),并推出面向农户的应用软件 FieldScripts,为肥力不均的地块提供最佳的作物品种选择和播种方案.杜邦先锋于 2013 年与农机巨头约翰迪尔(John Deere)合作,推出具有相似功能的 Field360.

1.2　育种相关数据大小

在近百年育种家的努力下,育种相关数据正处

于不断增长,特别是新一代测序技术的发展,直接带来了育种数据爆炸式的增加,以国际核苷酸序列数据库 GenBank 为例,它以指数式增长,大概 14 个月总数据量翻 1 倍.

1.2.1 基于测序的分子数据

由于物种多样性的存在,植物的基因组大小不一.一般大田作物基因组都在 1 Gb 左右,水稻基因相对小些为 0.4 Gb,大豆油料等为 1.0 Gb,而玉米、小麦等要几个 Gb 以上.英国皇家园林丘园焦佐尔实验室的遗传学家伊利亚·雷特彻发现,重楼百合(*Paris japonica*)拥有世界最大基因组——150 Gb[6].随着水稻、玉米等作物基因组被测序,越来越多的作物基因组序列不断公开发表,这些最终序列和测序过程中产生的数据,无疑是庞大的.美国国家生物技术信息中心(NCBI)的基因银行是全世界基因组数据的储存中心,公开发表论文的各类基因组、表达组数据均存储于此.目前,该中心的数据储存量已经达到万亿级.GenBank 发布的最新版本——Release 197(2013 年 8 月),已经涵盖超过280 000 个物种,数据年增长率达到 45.1%[7].2010 年,欧洲生物信息研究所(EMBL-EBI)的序列数据条目搜索为 4 亿个记录,而在 2014 年,记录数目已经超过了 10 亿条[8].在中国,"3K"水稻基因组项目,收集全球 2 859份水稻品种,产生了将近 16 T 的数据量[9].

1.2.2 大田观测数据

大田作物生长环境特点非常显著,不可控、变数大.因此,在大田中环境监测数据的记录就尤其重要.表现型数据的特点是采集需要耗费大量人力物力和时间成本高,并且,群体大小和性状、采集数据的地点都相当有限,远没有达到获得基因型数据的高通量低成本水平.随着自动控制技术、计算机和其他信息技术、图像处理技术的发展,性状表现型数据的采集也得到了一定程度的发展,但还远远没有达到成熟的程度.以我国育、繁、推一体的大型种子公司为例,一般的试验点数目已达到 200 个点左右,每年新增数据1 000万个以上,照片 10 万张以上[10].

同时,伴随高通量表现型分型技术和高通量基因分型技术的迅猛发展,对海量数据进行采集、加工处理、分析统计、可视化和最终应用于育种决策过程已经成为时代的标志,换言之,精确育种正式迎来了大数据(big data)时代.对高通量数据的采集技术已然成为促使作物产量再创新高的热点研究.如何储存和处理图形数据,使其与测序数据融合,对研究将能带来更大的便利.

2 大数据技术利用

大数据因具备高度战略意义、可操作性和产生巨大商业价值,引起研究者们的普遍关注[11].美国咨询公司麦肯锡(McKinsey)于 2011 年 5 月发表著名研究报告"Big data: The next frontier for innovation, competition, and productivity",标志着大数据时代的到来[12].大数据的兴起,主要源于因特网、云技术和物联网的迅速发展,各种终端设备、传感器和检测装置无时无刻不在产生数据.由于产生的数据量膨胀,同时由于传统方法处理对象的局限性,寻求处理此种窘境的方法呼之欲出.

大数据有 5 大特征,即所谓 5V:数量巨大(volume)、类型多样(variety)、处理速度快(velocity)、价值密度低(value),真实性(veracity)[11,13-15].在这 5V 中,数量巨大,类型多样指数据量大而形式多样,同时要求处理速度要快,而其中价值密度低则指的是数据信息存在垃圾多、污染重以及利用难的问题,然而就是在这样的低密度中却实实在在蕴涵着巨大的价值[16].可以说,大数据时代的到来将对研究方式、思维方式乃至于生活方式和生产方式都产生革命性变化.

Wamba 等[11]对已经发表截止到 2012 年 12 月 27 日的1 153篇研究结果和文献报道进行统计,通过人工处理筛选了最具代表性的 62 篇文献进行数据分析,发现自 2008—2012 年关于大数据的文献呈显著增长状态(图 1).同时,Wamba 等[11]根据产业类型分类发现,大数据研究在技术概念和服务领域以具占 16%的比例领先,而在医疗领域及政府管理方面的文献报道也分别占 11%和 7%(表 1).

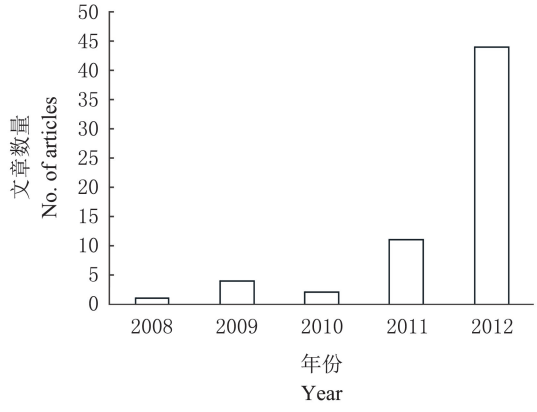


图 1 大数据相关文献(2008—2012 年)增长情况[11]

Fig.1 Distribution of articles by the year of publication[11]

表 1　大数据论文按产业类型分布[11]

Table 1　Classification based on industry[11]

应用领域 Application	文献数量 No. of articles	百分率 Percentage/%
零售业 Retail	3	6.7
医疗健康 Healthcare	5	11.1
生态 Ecology	1	2.2
教育 Education	2	4.4
政府 Government	3	6.7
制造业 Manufacturing	1	2.2
服务业 Services	7	15.6
技术领域 Technology	7	15.6
其他 Others	16	35.5
总数 Total	45	100.0

数据爆炸,直接面临的问题便是数据的存储.针对这一问题,科学家已研发了不同数据储存系统,包括 IBM 专家集成系统 PureData,浪潮云海大数据一体机,曙光 XData 大数据一体机,甲骨文 Exadata X3 一体机,EMC Greenplum 大数据一体机,华为一体机(王迪和范平,2013,中关村在线).Marx[17]建议面对膨胀的生物数据,云存储和云计算可以作为科学家应对这一状况的策略.同时,Spjuth 等[18]表示加大利用生物信息分析工作流程的利用,可以缓冲生物大数据的冲击.

对于大数据分析利用案例,最经典莫过于 Google 美国流感预测的应用.2009 年,Google 工程师在《Nature》科学期刊发表 1 篇名为《运用大数据的分析》的论文,该文解释 Google 如何利用民众搜寻的关键词,即能精确预测美国在冬天将爆发流感;果然那年冬天,美国发生 H1N1 流感危机,Google 系统所提供即时的资讯,有效协助公共卫生当局控制疫情[19].这是一个令世界瞩目的案例,让全世界意识到大数据不可估量的价值和应用前景.同时,Farecast 利用 10 万亿条航班价格记录,准确预测航程票价变化趋势,这同样引发人们对于大数据利用的重视[20].

近年来,美国已将大数据作为发展战略提上议程.自 2009 年 1 月 21 日现任美国总统奥巴马宣誓就职后的第一个工作日就签发“开放政府”备忘录,实施数字革命带动政府变革,到 2014 年 3 月,美国政府向全社会发出为政府大数据战略发展征询意见[21].

大数据科学应用发展的迅猛,对于人才的需求猛增.国外众多知名大学已纷纷加入大数据应用人才培养队伍,如美国斯坦福大学设立信息管理与分析专业,弗吉尼亚大学设立数据科学研究所等.在国内,针对大数据人才的培养方案则初露头角.2014 年,清华大学成立了清华-青岛数据科学研究院,以数据共享和整合为基础,研究应用为核心建立大数据分析共享平台(http://news.tsinghua.edu.cn/).2015 年 10 月,复旦大学成立了大数据学院和大数据研究院,将以计算机科学、数学和统计学为基础,与经济金融、生命科学、医疗卫生和社会管理等众多学科领域进行深度交叉研究,旨在有效推动相关学科的发展,直接面向产业需求建立跨学科、跨领域的研发团队,集聚产业创新人才(http://news.fudan.edu.cn/).

大数据的出现可以实现科学研究从过去的假设驱动型转化为数据驱动型,对大数据及相关处理技术可转化为巨大的社会经济价值,被誉为“未来的新石油”[22].

3　大数据技术在种业利用现状

现代遗传育种存在诸多问题,如长期育种选择导致遗传基础狭窄;杂交育种需时较长,近年发展的生物技术存在很多实践应用问题;公共平台和资源共享不够;传统育种方法和现代生物技术有待进一步融合等.现代育种技术体系复杂,需要多个学科交叉和多种技术支撑,缺乏有效数据组织和管理.分子设计育种作为新兴的育种技术体系,可以实现育种的定向变异、准确选择的目标.但是这一系统的应用,必须凭借强大的信息平台建设、共享机制以及计算模拟集成技术.在信息建设方面,现有品种信息库、核心种质信息库、重要性状基因功能与调控网络信息库、性状形成的生理生化信息库、分子标记数据库、生物信息学信息平台、生物统计分析平台等至关重要.但是我国育种领域数据孤岛与数据海洋问题严重.我国育种相关数据量很大,但分散,未有效组织.目前育种者在育种过程中利用的数据主要为自

身内部数据,而公开的文献和基因组相关数据等其他数据很少利用或无法利用.导致大量内部数据成为"数据孤岛",同时大量公开的育种相关数据(如基因组数据)成为"数据海洋",无从下手.上述问题极大限制了育种相关数据的利用和育种效率的提高.

目前,已有一些组织和部门构建了一些数据库和共享平台,但这些数据库往往存在相对分散、整合度不够高、针对性不够强等问题.如何有效利用科学家多年来产生的育种数据,整合处理成有效资源,并反哺于农业发展,大数据育种系统的发展,将会成为解决悬而未决育种难题的有效手段.

大数据育种策略不同于以前提出的一些计算机辅助育种方法,后者通常是将通用性状和特征性状进行数字化,便于利用遗传算法进行统计和筛选[23].王建康等[24]提出"植物育种模拟方法旨在建立较真实的遗传模型,对育种程序中的各种因素进行模拟筛选和优化,提出最佳亲本选配和后代选择策略.模拟方法利用经典遗传学、数量遗传学和群体遗传学的基本原理,结合各种遗传研究结果,定义育种性状的遗传模型".模拟育种的前提是基于特定子代性状来源于亲本,所以育种家们通常通过一些模拟方法评价亲本是否合适及亲本的影响程度,这些方法包括评估杂种优势表型表达状态,最佳线性无偏预测,基于谱系、分子标记数据的遗传关系分析等[25-34].王建康等[24]提出一个遗传模型所包含控制性状的基因有多少、它们在染色体的位点、每个基因座位上的等位基因数、等位基因间的作用方式及不同座位上基因间的作用方式等内容.育种模拟的优势在于可以比较不同标记辅助选择方法的育种效率(田间试验需要很大时间人力资源成本),提供有效遗传信息.但是,模拟育种只是基于积累的遗传数据和已知遗传原理进行虚拟育种设计,成功率尚待商榷.而遗传数据一经特定模型算法的限制,往往会造成信息丢失.

3.1 国外大数据技术在育种中的应用现状

先锋、先正达、孟山都等国际知名育种公司,都分别建立了高水平的育种信息平台和育种体系,在他们的研发队伍中,除了传统育种队伍和分子监测与分析队伍,都配备了一支庞大的生物信息和数量遗传学分析队伍.他们的育种工作人员除了在田间进行育种工作外,也会使用先进的数据采集设备,其育种数据库,包含了详尽的系谱信息和亲缘关系.这些育种公司能在育种行业处于垄断地位,与其大量采集处理的育种数据有着密不可分的关系.

在政府层面上,大数据技术利用往往是目前一个重点扶持产业之一.美国国家健康中心于2010年将1 000个基因组计划项目的数据上传至亚马逊云计算平台,研究者在使用其约2 700个体数据记录的同时也可以在该平台上传和储存数据[35].在2012年美国公布的大数据研究与发展方案中,即提出84个计划,范围涵盖国防、医疗、教育、能源、交通运输、国土安全、商业、科学、工业等应用领域[21].为响应创新领域发展,Intel与MIT等顶尖大学结盟成立,投入大数据核心技术开发;产业界IBM及GE,也专注发展特定领域之大数据应用.同时,美国政府创建了Data.gov网站,率先公开白宫相关数据,以期实现大数据的共享时代[36].最近有报道,美国伊利诺斯州的国家超级计算机应用中心获得180万美元的项目资助,用于大数据育种的开发(http://www.eurekalert.org/pub_releases/2015-07/crwi-ir071315.php).欧洲方面,Oxford的大数据中心专注在药物的开发.亚太地区的新加坡政府与GE已共同成立,协助企业发展大数据应用.欧亚地区还有很多国家(如英国、印度)已然开始了"数据公开"运动.总体上看,目前国外政府导向的大数据育种中研究与应用尚在起步中.

3.2 国内大数据技术在育种中的应用现状

在国家政策上,尽管政府相关部门已经将大数据技术与应用提上议程,但相较于西方国家,我国仍缺少政策型规划和经费建设.特别在大数据科学研究和技术开发层面,国家自然科学基金委员会第89期双清论坛"大数据技术与应用中的挑战性科学问题"上指出,我国科技界因缺乏经费和数据资源的支持使其积极性和主动性表现不高[22].我国大数据产业面临的问题包括数据开放共享程度低,数据安全和隐私保护风险日益突出,技术创新与应用能力滞后,产业生态体系尚未完善等[37].

我国工信部与大学研究机构,联手成立大数据研究中心.我国2010年起筹划构建的国家农业科学数据共享中心,主要收集从事农业科技活动所产生的基本数据,以及按照不同需求而系统加工整理的数据产品和相关信息,从而致力于推动科技资源优化配置,实现开放共享[38].但是该平台由于没有大数据的整合设计,整个农业科技平台的运作还停留在小数据时代.

我国已开始部分育种相关数据的研究和项目设计.如近期北京市科委重大项目"作物育种大数据技术与性状采集智能装备的研发与应用"也已经启动;

深圳市扶持华大基因的生物育种,建设农业大数据计算平台等.一些研究者已开始意识到大数据的重要性,叶锡君等[36]建议对水稻、大豆等十几种主要作物的创新种质、遗传材料、代表性地方品种等特种遗传资源进行数字化、标准化整理,构建农作物特种遗传资源共享平台[39].

4　大数据技术育种利用途径与展望

4.1　大数据技术育种利用途径

当前作物育种领域一个重要命题是如何总结和凝练大数据环境下农作物育种领域的创新方法,为我国"十三五"科技重点专项育种技术创新提供方法支撑.为此,本文提出一个大数据技术为基础的育种方法创新方案:基于知识工程的流程,有机整合作物不同品系、野生资源等在材料、基因、性状等方面的数据库,消除数据孤岛,形成大数据下的作物育种数据库及其处理系统;以性状数据采集和处理分析为核心,以作物育种过程管理为基础,研究作物育种资源整合、数据科学分析、过程信息化管理的育种技术新体系(图 2).

在涵盖农业育种数据信息的大规模数据库平台和富集算法及生物数字模型的计算体系基础上,以创新方法应用为契机,通过基因挖掘和育种技术,形成以用户需求(育种家或者企业)为导向的大数据育种技术平台.同时,创造有重大应用价值的新种质,培育和应用一批具有市场竞争力的突破性重大新品种,实现种质创新,提升育种自主创新能力.

图 2 表明,一个理想的大数据育种技术是以生产特定品种为导向,通过遗传信息流平台建设,根据需求组合信息流,输出满足目标基因组成的新品种.其中具体实施步骤包括:

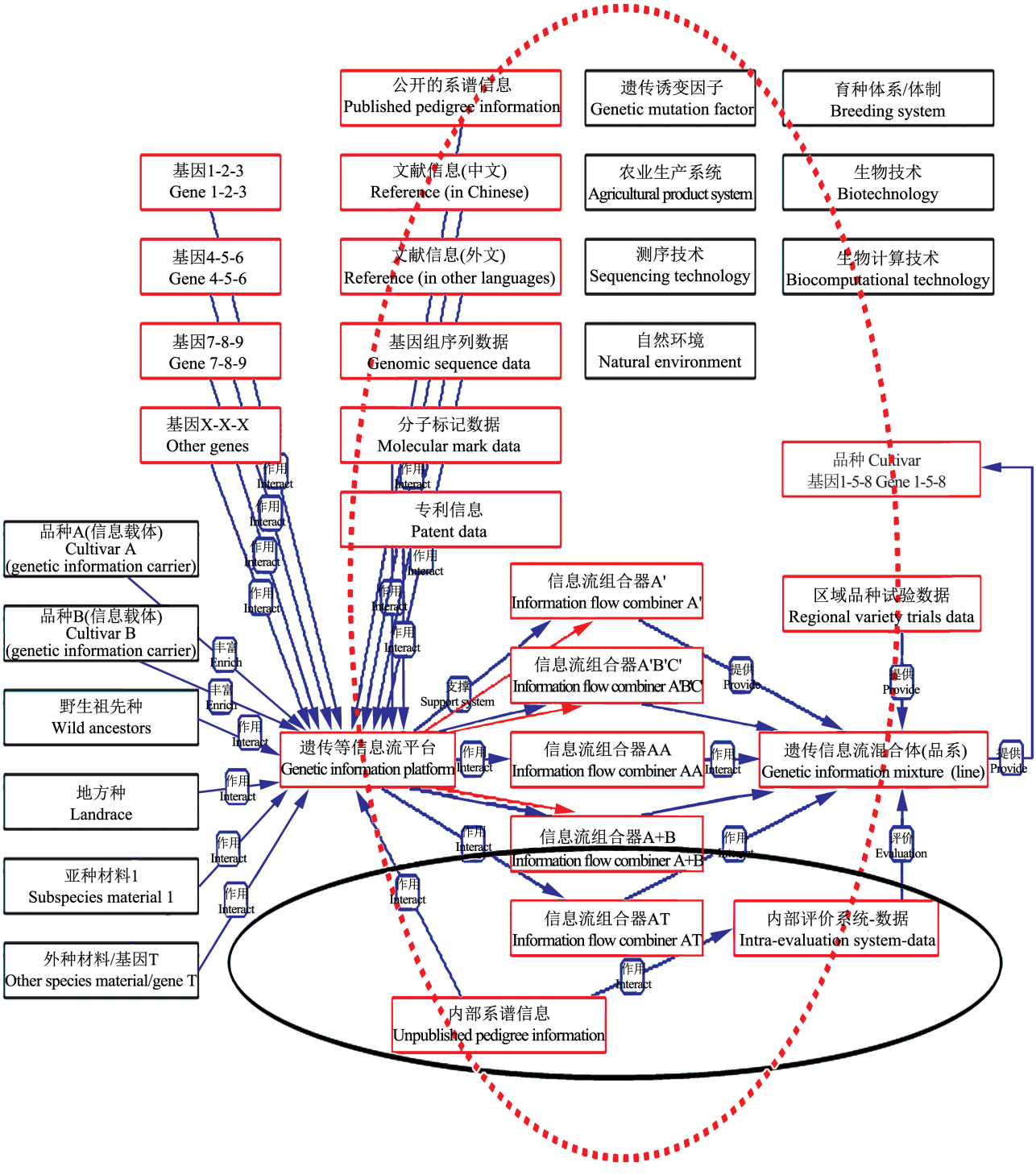
1) 收集遗传作物育种相关数据.获得育种与遗传材料相关文献(论文、育种相关书籍、专利等)、遗传资源相关数据、品种审定相关数据(品种区域试验数据和品质、抗性等测定数据等)和基因组相关数据(基因组、分子标记、基因序列、基因表达等公开数据).同时,各个育种组内部数据(系谱和田间表型和室内考种数据;分子数据等)也可以有效地采集并作为大数据系统的一部分加以利用.大数据育种最重要的基础是获得完整的育种相关数据,这是进行大数据育种技术的先决条件.

2)处理育种相关数据,进行统计分析,建立数据挖掘平台.如支持向量机、神经网络、序列模式发现等多种数据挖掘算法,均可以用于育种相关数据分析;同时数据运算采用云技术等,保证分析快速完成,及时提供分析结果.该步骤是利用大数据技术挖掘育种相关大数据形成概念/知识/育种建议的过程.

3) 搭建人机交互系统.以育种相关数据库及其育种信息与咨询系统形式出现,它不同于传统意义上的育种技术(如杂交育种、杂种优势育种技术等),但可以预计该技术将对作物育种工作产生巨大影响和作用.平台可以提供多元化育种服务内容,除了大数据挖掘与育种利用分析服务外,可以整合目前已有一些计算机辅助育种系统,如育种和实验数据辅助系统、田间设计与统计系统等.

4.2　大数据技术育种建议与展望

大规模生产、分享和应用数据的时代正在开启.数据的积累可以从量变引发质变,越来越多的企业、行业和国家以数据为资源进行知识和智力开发,挖掘数据价值,已经初具大数据思维[40].2010 年,美国科学家 Holdren 等呼吁集合国家之力解决各个研究领域出现的数据膨胀问题,并且为大数据技术加大财政支持[35].如近期美国伊利诺斯州超级计算机应用中心刚得到项目资助.我国目前在下一个 5 年(2016—2020)重大科技专项指南中也涵盖了大数据在育种领域应用的内容.为此,我们提出如下建议:1)开展我国作物育种相关数据本底和规模调查与估计.目前育种相关数据主要包括各个育种组内部数据、文献、遗传资源相关数据、品种审定相关数据、基因组相关数据等.应对这些数据规模、分布等进行全面调查,特别是育种课题组内部数据规模(如可分南方水稻和北方小麦)进行调查;2)开展基于大数据技术框架的育种相关数据采集、整合、挖掘与育种利用技术研究.作物育种相关数据,特别是表型数据采集是育种过程的一个难点,消耗大量人力和物力,数据准确性低误差大;同时,文献和相关数据等大量育种相关数据分散在各个数据库或期刊书籍中.结合计算机抓取和图像识别等技术,研发育种相关数据大规模和自动采集技术.开展育种相关大数据整合和挖掘技术,同时开展在育种中利用途径研究;3)开展大数据育种机构建设.建议成立全国性大数据作物育种中心,可考虑以北方和南方地域布局,也可以按照不同作物种类(如水稻/小麦/玉米等)建设.这些机构将对我国育种相关数据进行采集、归类和数据挖掘,同时提供公开的大数据育种平台分析与育种利用服务.



实线圈:现育种技术,主要利用育种小组内部数据.虚线圈:大数据育种技术,利用所有育种相关表型和分子数据,包括内部数据.

Solid line circle: data used by individual breeding program, including many private data. Dashed line circle: data used by the big data-based breeding technology, which include all data such as phenotype and molecular sequence from public resources and individual breeding programs.

图 2 大数据作物育种技术流程概念图

Fig. 2 Conceptual framework of big data-based crop breeding approach

农作物育种领域有着丰富的种质资源、海量各类型育种相关数据、漫长的育种过程及其复杂的技术系统,使得农业育种已然隶属大数据领域,构建大数据育种系统势在必行.历史上中国的育种技术曾领先世界,而在近现代的科学技术革命中,中国则退居学习者或跟踪者地位.这次大数据技术浪潮,为农作物育种变革提供了良机,是中国缩短与世界育种水平距离的机会,我们当以创新的魄力和勇气去抓住此次时代赋予中国的机遇.

参考文献(References):

[1] 董春水,才卓.现代数字育种技术的研究进展.玉米科学,2013,21(1):1-8.
Dong C S, Cai Z. Advanced in modern data-driven breeding technologies. *Journal of Maize Sciences*, 2013,21(1):1-8. (in Chinese with English abstract)

[2] 滕海涛.数字化玉米育种思路.中国农学通报,2008,12(24):495-498.
Teng H T. Exploration on digital maize breeding. *Chinese Agricultural Science Bulletin*, 2008,12(24):495-498. (in Chinese with English abstract)

[3] 孙健冬.高通量测序技术在农作物全基因组序列测定中的应用概览.生物技术进展,2012,2(1):11-15.
Sun J D. A brief review of crop's whole genome sequencing by next generation sequencing technology. *Current Biotechnology*, 2012,2(1):11-15. (in Chinese with English abstract)

[4] 王冰冰.大数据:植物育种的加速器.高科技与产业化,2015,5(228):50-52.
Wang B B. Big data: the accelerator of crop breeding. *High Technology and Industrialization*, 2015,5(228):50-52. (in Chinese)

[5] Krajewski P, Chen D, Ćwiek H, *et al*. Towards recommendations for metadata and data handling in plant phenotyping. *Journal of Experimental Botany*, 2015, 66(18):5417-5427.

[6] Pellicer J, Fay M, Leitch I, *et al*. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 2010,164:10-15.

[7] Benson D, Clark K, Karsch-Mizrachi I, *et al*. GenBank. *Nucleic Acids Research*, 2014,42:D32-D37.

[8] Squizzato S, Park Y, Buso N, *et al*. The EBI Search engine: Providing search and retrieval functionality for biological data from EMBL-EBI. *Nucleic Acids Research*, 2015, 4:8.

[9] 郑天清,余泓,张洪亮,等.水稻功能基因组育种数据库(RFGB):3K 水稻 SNP 与 InDel 子数据库.科学通报,2015,4(60):367-371.
Zheng T Q, Yu H, Zhang H L, *et al*. Rice functional genomics and breeding database (RFGB): 3K-rice SNP and InDel sub-database. *Science Bulletin*, 2015,4(60):367-371. (in Chinese with English abstract)

[10] 王虎,杨耀华,李绍明,等.基于移动端作物大田测试数据采集技术研究与实现.中国农业科技导报,2013,15(4):156-162.
Wang H, Yang Y H, Li S M, *et al*. Research and implementation of field crop test data collection based on mobile phone. *Journal of Agricultural Science and Technology*, 2013, 15(4):156-162. (in Chinese with English abstract)

[11] Wamba S, Akter S, Edwards A, *et al*. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 2015,165:234-246.

[12] Manyika J, Chui M, Brown B, *et al*. Big data: The next frontier for innovation, competition, and productivity, 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[13] Russom P. The three Vs of big data analytics, *TDWI*. 2011.

[14] McAfee A, Brynjolfsson E. Big data: The management revolution. *Harvard Business Revolution*, 2012,10:61-68.

[15] Kwon O, Sim J. Effects of data set features on the performances of classification algorithms. *Expert System Application*, 2013,40(5):1847-1857.

[16] White M. Digital workplaces: Vision and reality. *Business Information Review*, 2012,29(4),205-214.

[17] Marx V. The big challenges of big data. *Nature*, 2013,498(6):255-260.

[18] Spjuth O, Bongcam-Rudloff E, Hernandez G, *et al*. Experiences with workflows for automating data-intensive bioinformatics. *Biology Direct*, 2015,10:43.

[19] 涂子沛.大数据:正在到来的数据革命,以及它如何改变政府、商业与我们的生活.桂林:广西师范大学出版社,2013:1,10.
Xu Z P. *Big Data: Incoming Data Revolution, and How it Changes Governments, Business and Our Lives*. Guilin: Guangxi Normal University Press, 2013:1,10. (in Chinese)

[20] 维克托·迈尔-舍恩伯格,肯尼思·库克耶.大数据时代:生活、工作与思维的大变革.盛杨燕,周涛,译.杭州:浙江人民出版社,2012:58.
Mayer-Schöenberger V, Cukier K. *Big Data: A Revolution that Will Transform How We Live, Work and Think*. Cheng Y Y, Zhou T, translate. Hangzhou: Zhejiang Public Press, 2012:58. (in Chinese)

[21] 大数据的国家战略.信息系统工程.新闻透视,2015,4:8-9.
Big data of national strategy. Information System Engineering. *News Insight*, 2015,4:8-9. (in Chinese)

[22] 王成红,陈伟能,张军,等.大数据技术与应用中的挑战性科学问题.中国科学基金,2014(2):92-97.
Wang C H, Chen W N, Zhang J, *et al*. Challenging scientific problems for technologies and applications of big data. *Science Foundation in China*, 2014(2):92-97. (in Chinese with English abstract)

[23] Azimzadeh M, Amiri R, Davoodi-Bojd E, *et al*. Computer aided selection in breeding programs using genetic. *Spanish*

Journal of Agricultural Research, 2010, 8(3): 672-678.

[24] 王建康,李慧慧,张鲁燕.基因定位与育种设计.北京:科学出版社,2014,6:168,230.
Wang J K, Li H H, Zhang L Y, *et al*. Gene Location and the Design of Breeding. Beijing: Science Press, 2014, 6:168, 230. (in Chinese)

[25] Melchinger A, Schmidt W, Geiger H. Comparison of testcrosses produced from F_2 and first backcross populations in maize. *Crop Science*, 1988, 28:743-749.

[26] Dudley J. Breeding: Choice of parents//Goodman R M(ed). *Encyclopedia of Plant and Crop Science*. London: Taylor & Francis, 2004:215-217.

[27] Panter D, Allen F. Using best linear unbiased predictions to enhance breeding for yield in soybean: Ⅰ. Choosing parents. *Crop Science*, 1995, 35:397-405.

[28] Burkhamer R, Lanning S, Martens R, *et al*. Predicting progeny variance from parental divergence in hard red spring wheat. *Crop Science*, 1998, 38:243-248.

[29] Bernardo R. *Breeding for Quantitative Traits in Plants*. Woodbury, MN, USA: Stemma Press, 2002.

[30] Dudley J, Maroof M, Rufener G. Molecular marker information and selection of parents in corn breeding programs. *Crop Science*, 1992, 32:301-304.

[31] Bernardo R, Yu J. Prospects for genome wide selection for quantitative traits in maize. *Crop Science*, 2007, 47: 1082-1090.

[32] Zhong S, Jannink J. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics*, 2007, 177:567-576.

[33] Frisch M, Thiemann A, Fu J, *et al*. Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theoretical Applied Genetics*, 2010, 120:441-450.

[34] Sun X, Peng T, Mumm R. The role and basics of computer simulation in support of critical decisions in plant breeding. *Molecular Breeding*, 2011, 28:421-436.

[35] Mervis J. Agencies rally to tackle big data. *Science*, 2012, 336:22.

[36] 侯人华,徐少同.美国政府开放数据的管理和利用分析:以www.data.gov为例.图书情报工作,2011,4:119-122.
Hou R H, Xu S T. Management and Reuse of the US. Government open data: Taking www. data. gov for an example. *Library and Information Service*, 2011, 4: 119-122. (in Chinese)

[37] 王伟玲.大数据产业的战略价值研究与思考.技术经济与管理研究,2015,2:117-120.
Wang W L. Research and thinking on the strategic value of the big data industry. *Technological Economy and Management Research*, 2015, 2: 117-120. (in Chinese with English abstract)

[38] 曹永生,方沩.国家农作物种质资源平台的建立和应用.生物多样性,2011,18(5):454-460.
Cao Y S, Fang W. Establishment and application of national crop germplasm resources infrastructure in China. *Biodiversity Science*, 2011, 18(5):454-460. (in Chinese with English abstract)

[39] 叶锡君,孙敬,张天真.农作物特种遗传资源共享平台的建立.南京农业大学学报,2011,34(6):7-12.
Ye X J, Sun J, Zhang T Z. Genetic resources sharing platform construction of crops. *Journal of Nanjing Agricultural University*, 2011, 34(6):7-12. (in Chinese with English abstract)

[40] 张维明,唐九阳.大数据思维.指挥信息系统与技术, 2015, 6(2):1-4.
Zhang W M, Tang J Y. Big data thinking. *Command Information System and Technology*, 2015, 6(2): 1-4. (in Chinese with English abstract)