

林木大数据与智慧育种的机遇与挑战

钮世辉 李 悦*

(北京林业大学林木资源高效生产全国重点实验室, 北京 100083)

摘 要: 与历经数千年历史的作物育种相比, 现代林木育种历史不足百年, 其理论与技术多处于跟踪状态, 而随着现代分子生物学、生物技术、生物组学、生物信息学和人工智能等技术迅速发展及其在林木遗传育种中的应用探索, 林木遗传育种与作物和园艺植物育种的差距正在逐渐缩小。当前, 以大数据和人工智能为标志的植物育种 5.0 概念被提出, 而林木育种是否可跨入到 5.0 时代, 其面临的机遇和挑战如何, 成为林木育种工作者感兴趣的新议题。

主要的人工林树种基本都是多年生的异花授粉植物, 应用场景以自然或人工改良的自然环境为主, 林木良种要有优异的经济性状表现, 还要适应一定的自然或栽培条件。树木的生殖和经济性状形成的周期较长, 经济性状为多基因调控, 杂合体亲本不能培育纯系良种, 轮回选择制约, 育种程序复杂, 遗传转化困难等问题, 成为限制林木育种理论与技术突破的瓶颈。而植物育种 5.0 的时代的到来, 是否能成为林木育种破局的机遇, 应该有所期待。

本研究将阐述“林木大数据与智慧育种”概念提出的时代背景, 概述植物育种技术的突破和对育种发展的贡献, 陈述现代林木育种技术发展简史及其跟踪特点与滞后现实, 提出理想的林木遗传改良技术发展愿景: 集成、精准、高效。

以大数据和人工智能为标志的植物育种 5.0 概念的提出, 推动了该发展方向的积极探索与实践, 但截至目前仍尚未有成功的模式和样板。林木育种 5.0 应该怎样做也是我们要探索的问题。林木育种 5.0 既然也要有大数据和人工智能的基本特征, 就不会是局限于树种和区域的小题杂议, 是要为大育种时代提供全新的理论与技术策略的战略高地, 要立足气候与需求变化趋势, 以共性为基础, 特性为引导, 为各区域的林木育种提供更优的理论、程序和技术措施的指导。即: “系统整合、共特有别、按需利导”。

林木育种 5.0 应该是具有战略意义的宏大科技创新体系及系统工程, 要有泛国际、国内、区域或树种创新联盟的广泛共识、会聚与合作, 有健全的组织体系与科学、民主、公平、共享、高效的运行机制保障, 有来自政府、基金和企业事业机构的人力与资源投入, 有执行过程的监管和绩效评估体系保障, 有科技成果的有效实现途径。

大数据是林木育种 5.0 的基石。各种变异信息是林木遗传育种数据的基本属性。主要包括在遗传测定基础上的遗传单元各性状的表现型变异信息、遗传型变异信息、环境因子变异信息、遗传型与环境互作变异信息。传统遗传育种信息为具体经济性状、遗传来源变异、加性与非加性遗传效应、气象或立地综合效应等, 只能解决特定群体与性状的变异规律问题, 为育种策略和增益预测提供经验依据。并积累了大量的同质性信息。但基于有限数据与树种特点的局限, 对变异共性规律和特性的形成缺乏系统整合与鉴别, 未能形成对育种性状变异的因果关联机制的系统性认识体系, 高效育种缺乏更深刻理论依托, 育种策略仍受制于基于传统经验的多群体、长周期的低效轮回选择体系。

林木育种 5.0 时代的大数据应该是超越时空、学科、技术、环境、树种、性状、指标等所有深层次信

息来源的系统性会聚、整合、分类、综合、纵横与解析。就性状表型变异而言,性状就不局限于形态、生长、发育和适应性的外在状况,导致这些表型变异的内在解剖学、细胞学、生理、生化、次生代谢等指标信息均要列入大数据范畴,而这些信息的获得需要有高效与精准的测定技术支持。而导致表型变异的遗传型变异大数据则不仅要有遗传单元的种群遗传结构和亲缘关系信息,还要有决定变异的表型组、基因组、调控组、转录组、蛋白组、代谢组、生理组、生化组、生态组等模块信息,模块间的互作网络与关键节点等对决定性状变异有重要意义;环境型大数据主要有光环境、大气环境、立地与土壤环境、有益及有害生物环境、人工林培育技术环境构成,自然与栽培条件的综合环境信息是基础,而在拟态试验下的单一环境因子对表型的决定性影响有更大的价值。而可以预见遗传型与光、气、地、生、人各环境因子模块的单一或综合互作形成的光效表型组、气效表型组、地效表型组、生效表型组、技效表型组等,对我们深化林木经济性状变异的形成机制,构建全新的功效育种理论和技术策略,有着令人向往的期待。

基于大数据的智慧林木育种,需要借助迅速发展的人工智能体系,构建具有较强深度学习能力且密切结合林木生物学特性的专业化信息解析与智慧判别体系,基于大数据提供的林木深层次的表型、遗传、环境和互作模块信息,构建由表型现象到遗传规律、变异本质、因果联系、模块体系、技术途径判断、育种策略抉择、良种技效预测等的人工智能科学决策体系,而林木育种科技人员学术成果、技术经验、育种实践、种质资源等知识要经过信息化处理并整合到人工智能系统之中。因此未来的林木育种 5.0 应该是由大数据智慧育种创新平台主导、学术信息库为主体、遗传育种学家和科技人员提供的经验知识、专业人工智能体系提供科学决策支撑依据,共性理论与特色需求协同的共享性智能化集成育种平台。在该体系中,高质量和大规模的数据与信息采集和积累是基础。大数据和人工智能介导的智慧育种体系将通过机器学习、优化和模拟育种程序,开展育种设计和选择结果预测,从而提高育种效率、改进遗传增益。

平台除了大数据信息智慧处理外,还应具备较强的各环境因子拟态试验技术平台,表型、遗传、环境等多学科层次变异信息的综合、高效、精准检测技术能力,提供更可靠的遗传与技术效果信息,结合大数据的发掘和解析,开展模拟、预测、优化、提供育种策略与技术的选择依据,实现林木育种项目的智能化,并为育种与基因资源的优化管理和品种权保护提供技术支撑。

提出的林木育种 5.0 设想是理想化的,要实现起来任重道远。首先可利用的林木育种大数据具有局限性,表型、遗传、环境、互作的信息整合具有一定难度,各树种的育种信息存在较大差别,信息欠充分,数据来源系统的可比性存在问题,数据的真实可靠性待验证,信息的科学有效性,关联性、可共享性都需要通过人工智能来判别、发掘和整合。林木智慧育种的人工智能系统尚未建立,系统的算法逻辑体系、不同信息模块的系统构建、系统间互作调控网络、高阶信息数据的会聚、解析,判别模型与算法构建,可靠性与绩效评判,理论模块与技术程序的精粹优化等均需要做原创性的探索与实践。