

doi: 10.3969/j.issn.1674-1242.2023.04.001

融汇计算和生物，孕育创新与人才 ——数字生物学实验室成立 20 周年纪念

许东

(密苏里大学电气工程与计算机科学系, 哥伦比亚, 密苏里州, 美国 65211)

【摘要】在数字生物学实验室 (DBL) 20 年的发展历程中, 我们见证了生物信息学与计算生物学的迅猛进步, 并见证了实验室在科研成果、人才培养和科研合作等方面的显著成长。DBL 展现了在识别新机遇和拓展新领域方面的强大适应力, 开发了许多创新的算法、软件和信息系统, 为解决多种生物学和医学问题提供了有效的工具, 还在数据分析和建模方面做了大量工作。在这个重要的 20 年里程碑时刻, 我们回望过去所取得的成就, 并以此为基础, 满怀期待地展望未来, 庆祝我们团队的共同努力和不懈追求。

【关键词】生物信息学; 计算生物学; 人工智能; 深度学习; 人才培养; 国际合作

【中图分类号】TP39

【文献标志码】C

文章编号: 1674-1242 (2023) 04-0325-06

Integrating Computing and Biology to Nurture Innovation and Talents — Celebrating the 20th Anniversary of Digital Biology Laboratory

XU Dong

(Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA)

【Abstract】Over the past twenty years, the Digital Biology Lab (DBL) has witnessed rapid progress in bioinformatics and computational biology, and it has achieved significant milestones in research, education, and collaboration. The DBL has shown strong adaptability by identifying new opportunities and branching into emerging fields. It has developed numerous innovative algorithms, software packages, and information systems that provide effective tools for addressing a wide range of biological and medical challenges. In addition, the DBL is actively involved in extensive data analysis and modeling for biomedical studies. As we have reached this significant twenty-year milestone, we reflect on our past accomplishments and look forward to future endeavors, celebrating the collective efforts and continuous commitment of our team.

【Key words】Bioinformatics; Computational Biology; Artificial Intelligence; Deep Learning; Talent Training; International Collaboration

收稿日期: 2023-12-01。

作者简介: 许东, 教授, 美国密苏里大学电子工程和计算机科学系的校董会讲座教授, 同时担任 Christopher Bond 生命科学中心研究员。于 1987 年和 1990 年分别获得北京大学物理学本科与硕士学位, 于 1995 年在伊利诺伊大学获得物理学博士学位。在美国国家癌症研究所做了两年博士后, 此后在美国橡树岭国家实验室从事了 6 年研究工作。2003 年加入美国密苏里大学, 2007—2016 年担任该校计算机科学系的系主任, 2017—2020 年担任该校信息技术专业主任。发表科研论文 500 余篇, 是美国科学促进会会士, 美国医学和生物工程研究院会士和亚太人工智能协会会士。

1 实验室简介

密苏里大学哥伦比亚分校的数字生物学实验室 (Digital Biology Lab, DBL) 成立于 2003 年 8 月。自成立以来, DBL 始终致力于开发与应用前沿的算法和软件平台, 培养了大批生物信息学、机器学习及软件方面的人才。目前 DBL 有 2 位教授 (我和 Trupti Joshi)、6 名职员、20 余名博士研究生及十几名硕士研究生, 还有不少本科生、实习生和访问学者。团队成员主

要分布在 Christopher Bond 生命科学中心、医学院和最近成立的 NextGen 精准健康中心。DBL 作为生物信息学和计算生物学领域的一个研究和教育中心, 主要致力于三方面的工作 (见图 1): ①研发新颖的算法 (机器学习、图论、统计等方法); ②开发生物信息学平台 (工具网站和信息系统); ③将算法与平台和其他信息学资源应用于各种生物和医学问题的分析、预测、建模中。

数字生物学实验室 (DBL, <https://digbio.missouri.edu>)

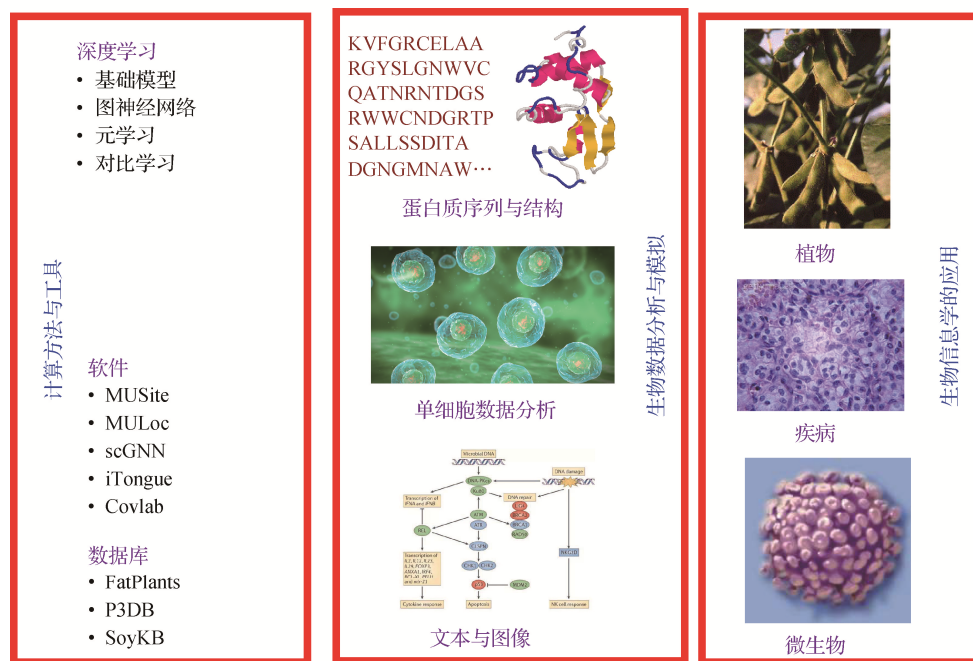


图 1 DBL 的主要科研方向

Fig.1 The main research directions of DBL

2 科研方向与成果

DBL 适应性强, 善于抓住新的机遇, 一直活跃在生物信息学的前沿领域。DBL 前十年的研究重点主要集中于蛋白质结构的预测与计算系统生物学, 近十年的研究重点则是生物信息学与深度学习的结合。通过开创性的深度学习及应用, DBL 取得了生物信息学方面的系列成功, 包括单细胞数据分析、蛋白质结构及属性的预测、基因型-表型关系、生物路径图例的图像理解, 以及舌像健康评估。这些应用利用了广泛的深度学习方法, 比以前的方法显著提高了性能。DBL 强调科研工作量的影响与质量, 确保文章具有新颖性、准确性、可复制性和清晰性, 这种严格的质量控制反映了 DBL 对学术诚信和研究质量的高度重视。

DBL 一直没有停止过蛋白质分析和结构预测的工作。DBL 开发了一种新的蛋白质结构预测系统——MUFOLD, 在全球蛋白质结构预测技术的评估中数次名列前茅^[1]。DBL 还研发了新颖的深度学习方法, 用于蛋白质分子动力学模拟分析^[2]、蛋白质定位预测、蛋白质翻译后修饰位点预测^[3]。目前 DBL 正在积极开发蛋白质序列与结构一体的基础模型, 并用提示学习的方法及蛋白质大模型进一步提高蛋白质属性预测的性能^[4]。

DBL 过去几年在单细胞数据分析方法开发方面做了一系列原创工作。特别值得一提的是, DBL 利用图神经网络成功设计了算法和工具来破译大规模、嘈杂的单细胞数据中涵盖的复杂生物模式与机制。例如,

DBL 的 scGNN 工作^[5]首次将图神经网络框架引入单细胞数据分析,得到了广泛的引用与跟进,发表仅两年就被引用了 300 多次。DBL 进一步将图神经网络框架用于分析空间转录组学及其他模态的数据,并且取得了很好的结果。DBL 在《自然·通讯》期刊上发表的 *Dimension-agnostic and granularity-based spatially variable gene identification using BSP* 一文^[6]入选了该期刊的编辑亮点,成为生物技术和方法领域最近发表的 50 篇最佳论文之一。目前 DBL 正在探索单细胞数据大模型的应用,并有了一定的进展。除此之外,DBL 正在开发“一站式”单细胞数据分析及方法测试平台。这些工作为理解发育和疾病中复杂组织的异质性及潜在机制提供了重要的手段。

近年来,DBL 积极探索生物学文献的挖掘及生物医学图像的解析。在文本数据挖掘方面,DBL 采用 ChatGPT 和其他先进的语言大模型来提取基因关系和生物学路径的重要信息。为了提高 ChatGPT 反馈的准确性,DBL 引入了创新的迭代提示化技术。DBL 运用 SAM 等基础模型对生物医学图像进行图像理解,从科研论文的示意图中识别实体及其关系。DBL 还将提示学习用于 SAM,以检测冷冻电子显微镜(cryo-EM)图像中的蛋白质。

DBL 注重生物信息学工具和数据库的研发,主导开发了 20 多个拥有众多用户的生物信息学软件工具和信息系统,服务了广大用户,推动了生物学和医学研究的深入。相关成果具体如下。①大豆知识库(SoyKB,大豆研究的综合网络资源)^[7];②P3DB(植物蛋白磷酸化数据存储库)^[8];③FatPlants(研究植物脂肪相关基因、蛋白质和代谢的数据与工具);④Musite 和 MusiteDeep(预测蛋白质翻译后修饰)^[9];⑤MuLoc 和 MuLocDeep(预测蛋白质定位)^[10];⑥MuFold(预测蛋白质结构);⑦G2PDeep(预测定量表型和基因组标记物);⑧scGNN(单细胞 RNA 测序分析);⑨BSP(预测空间可变基因);⑩VPE(使用患者检查数据的虚拟学习环境);⑪iTongue(基于舌头的健康诊断移动应用程序);⑫Covlab(基于推特数据预测新冠疫情趋势)。

DBL 在各种研究课题上与许多实验学家积极合作,利用生物信息学工具在生物学和医学研究中进行深入应用。相关计算与分析有助于实验设计,并为许

多研究中的生物学问题提供了新的见解。这些应用涵盖广泛的物种(如病毒、细菌、植物和人类)和领域(如基因组分析、DNA 甲基化、非编码 RNA 分析、基因表达分析、基因功能预测、蛋白质结构预测和分析、蛋白质组学、蛋白质磷酸化、代谢组学、单细胞分析等)。DBL 还参加了几项具有高影响力的研究,如大豆的全基因组测序^[11]。

3 人才培养

DBL 始终将人才培养视为核心任务。算上在实验室做过至少一个暑期或一个学期科研工作的人员,前后一共约有 300 人在 DBL 学习和工作过,包括职员、博士研究生、硕士研究生、本科生、实习生和访问学者,他们中有很多人在学术界和工业界取得了可喜的成就。在学术单位从事科研工作的有 60 多人。一些青年学者已经成为相关行业的领军人物,并培养了下一代科研人员。也有人去了大型企业或正处于上升期的创业公司。在 DBL 做过科研的本科生、实习生中,有不少进入名校读研深造。

在人才培养方面,DBL 非常重视因材施教。DBL 不要求学生拥有特定的培训基础,唯一的要求是要有强烈的学习和跨学科研究动机。团队成员在加入团队之初,DBL 会和他们一起讨论职业目标。一旦明确了职业目标,DBL 就会为他们制订符合其职业发展的计划,并根据该计划对他们进行个性化培养。在科研选题上,DBL 会充分考虑他们的职业需求。对于那些希望在学术界发展的团队成员,DBL 会对他们的独立思考能力、表达能力要求更高,并多鼓励他们在期刊、会议上投稿。而对于那些想在工业界发展的团队成员,DBL 则会重点培养他们的实践操作能力和项目管理技能,并鼓励他们多做暑期实习,以帮助他们未来在企业中独当一面。

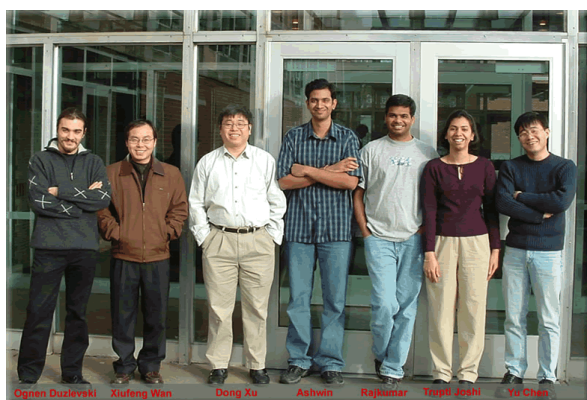
教师的基本职责是授业和解惑。授业是指向学生传授知识,如课堂教学,开展各种类型的辅导;解惑则是帮助学生解决在学习和科研过程中遇到的各种问题,满足学生的求知欲。然而,对从事科研工作的人来说,除了这两点,教师还有两项更加重要的职责。一是指明科研方向和选择合适的研究题目。在科研过程中,确定正确的方向和有价值的题目至关重要。有人说,一旦科研问题定义清楚了,问题就解决了一半。二是保证科研质量,不断给予学生点评与指导。在科

研讨论和文章撰写过程中,反馈与修改就像增强学习一样,教师通过不断指出问题和提出改进意见,帮助学生提高科研能力和科研质量。以上四个方面共同构成了教师在科研领域的重要职责,DBL 的教师和资深科研人员都积极参与了这个过程。

DBL 的培养模式注重个性化指导与团队合作。就我个人来说,我经常和我直接指导的每个成员进行各种交流,我每周都会有两三天时间去实验室和每个成员对话。团队成员们也经常约我长谈,我基本上没有拒绝过,只是希望他们每次讨论课题时用图表来展示数据结果,不能空谈想法。另外,就像非洲的一句谚语:“It takes a village to raise a child. (养育一个孩子需要整个村庄的帮助。)”,培养一名科研人才需要来自各方的支持和指导,确保被培养者能从多位资深科研人员与合作者那里获得全面的帮助。很多组员还接受其他老师的指导,如委员会成员、合

作者及其他任课老师。DBL 还设有固定的每周组会,很多项目也设有例会,组员之间会进行很多讨论。DBL 对组员的帮助一直持续到他们离开实验室,并在他们毕业后继续为他们提供职业建议和支持。

为了创造一个理想的工作环境,DBL 重视发展友好、愉快的交流环境。DBL 认识到,工作与生活的平衡对促进创新和卓越研究至关重要。为此,DBL 采取了多种措施。例如,每周的组会不仅是学术交流的平台,还提供午餐,组会结束后还安排了交流时间。此外,DBL 还举行茶歇和社交活动,如每学期的户外出游和聚餐,以增强团队成员之间的联系。我曾在家中举办各种聚会,为大家提供了丰富的交流机会,使工作氛围更加轻松愉快。这些活动不仅为团队成员带来了乐趣,也有效缓解了他们的压力,给他们留下了美好的回忆。DBL 成员的相关活动合影如图 2 所示。



(a) 2003 年, DBL 成员在密苏里大学的合影



(b) 2013 年, DBL 成员在 Lake of the Ozarks 10 周年庆典的合影



(c) 2018 年, DBL 在堪萨斯城 Stowers Institute 组织 15 周年庆典暨生物信息学论坛, 参会人员的合影



(d) 2023 年, DBL 在密苏里州哥伦比亚市组织 20 周年庆典暨生物信息学与机器学习论坛, 参会人员的合影

图 2 DBL 团队合影

Fig.2 DBL photos

4 合作与交流

作为一门交叉学科,生物信息学的边界不断扩展,其影响已经渗透到生物医学的各个分支领域,DBL 的研究也不断挑战和扩展现有技术与知识的边界。DBL 的软件和工具也被越来越多的研究所采用,加速了一些科学发现的步伐。所有这些都依赖紧密的科研合作,只有通过跨学科的合作,才能实现更有影响的成果。DBL 效仿卡文迪许实验室(Cavendish Laboratory),该实验室隶属英国剑桥大学,成立于 1874 年,以其开放和合作的学术氛围著称。该实验室鼓励跨学科合作,促进了各领域之间的思想交流和技术融合,培养了大量物理学领域的专业人才。DBL 虽然不能与卡文迪许实验室同日而语,但也是一个非常开放的实验室,吸引了各种背景的人才前来学习、工作与交流,是一个思想、文化、科学与技术的大熔炉。

在过去 20 年,DBL 与全球数十家实验室在不同时期建立了牢固的合作关系,不仅促进了问题的深入研究,也彼此取长补短,使科研工作再上一个台阶。尤其是中美两国之间的合作,让访问学者的科研水平得到了显著提升,也大大拓展了 DBL 科研工作的深度与广度。加上国内来访的老师、学生与“海龟”成员,现在 DBL 有 40 多人在国内拥有教职。不少国内来访的成员做出了非常优秀的科研工作,科研成果发表在高影响力的杂志上。

5 展望未来

20 年前的今天,DBL 在生物信息学和计算生物学领域扬帆起航,如今已经成长为该领域被业内认可的研究和教育基地。站在 20 周年的新起点,DBL 将继续拓宽研究视野,深化科研内容,不断推出更多高质量的论文和工具。DBL 还将继续致力于培养更多的科研人才,为全球科学进步贡献力量,同时期待与更多伙伴合作,共同探索生物信息学的未来。

2023 年 9 月,DBL 举办了 20 周年庆暨生物信息学与机器学习论坛,吸引了不少从前的成员。会议上大家不仅交流了各自在生物信息学和机器学习领域的开创性研究,还播放了回顾视频,展示了 DBL 的过往

和珍贵回忆。为庆祝 DBL 成立 20 周年,促进 DBL 成员的交流及分享,同时推动生物信息学及人工智能在生物医疗领域的应用和发展,《生物医学工程学进展》隆重推出了“DBL 20 周年庆”特刊,展示了不少 DBL 成员的优秀工作成果。在此,我真心感谢特刊的编委与作者,同时向现在和过去的成员以及所有支持、信任和参与 DBL 成长的合作伙伴表示最深的谢意!是你们的努力与奉献,成就了 DBL 20 年的辉煌。未来,我们将携手共创数字生物学更加灿烂的明天。

参考文献

- [1] ZHANG Jingfen, WANG Qingguo, *et al.* MUFOLD: a new solution for protein 3D structure prediction[J]. **Proteins: Structure, Function, and Bioinformatics**, 2010, 78:1137-1152.
- [2] ZHU Jingxuan, WANG Juexin, HAN Weiwei, *et al.* Neural relational inference to learn long-range allosteric interactions in proteins from molecular dynamics simulations[J]. **Nature Communications**, 2022, 13: 1661.
- [3] WANG Duolin, ZENG Shuai, XU Chunhui, *et al.* MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction[J]. **Bioinformatics**, 2017, 33(24): 3909-3916.
- [4] TANG Bowen, HE Fengming, LIU Dongpeng, *et al.* AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2[J]. **Biomolecules**, 2022, 12: 746.
- [5] WANG Juexin, MA Anjun, CHANG Yuzhou, *et al.* scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses[J]. **Nature Communications**, 2021, 12: 1882.
- [6] WANG Juexin, LI Jinpu, KRAMER S T, *et al.* Dimension-agnostic and granularity-based spatially variable gene identification using BSP[J]. **Nature Communications**, 2023, 14: 7367.
- [7] JOSHI T, MICHAEL R, *et al.* Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding[J]. **Nucleic Acids Research**, 2014, 42(1): D1245-D1252.
- [8] YAO Qiuming, GE Huangyi, WU Shangquan, *et al.* P3DB 3.0: from plant phosphorylation sites to protein networks[J]. **Nucleic Acids Research**, 2014, 42(1):D1206-D1213.

- [9] WANG Duolin, ZENG Shuai, XU Chunhui, *et al.* MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction[J]. **Bioinformatics**, 2017, 33(24): 3909-3916.
- [10] JIANG Yuexu, JIANG Lei, *et al.* MULocDeep web service for protein localization prediction and visualization at subcellular and suborganellar levels[J]. **Nucleic Acids Research**, 2023, 51: W343-W349.
- [11] SCHMUTZ J, STEVEN B, SCHLUETER J, *et al.* Genome sequence of the palaeopolyploid soybean[J]. **Nature**, 2010, 463:178-183.