大数据时代的转化生物医学信息学*

白晋伟 夏开建

（苏州大学图书馆数字化部 苏州 215006） （苏州大学附属常熟医院 常熟 215500）

钱福良 姜 智 朱 斐 沈百荣

（苏州大学系统生物学研究中心 苏州 215006）

〔摘要〕 转化生物医学信息学是生物信息学、医学信息学与转化医学多学科融合的新型学科，讨论该学科的兴起、学科生态，指出充分的数据共享是基础、完备的基因型－临床表型数据是核心、解决复杂疾病的精准预测问题是关键，深入分析该学科的内涵及大数据背景下发展趋势。

〔关键词〕 生物信息学；医学信息学；转化医学；生物医学大数据

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.01.001

Translational Biomedical Informatics in the Era of Big Data *BAI Jin－wei, Digital Department of Library, Soochow University, Suzhou 215006, China; XIA Kai－jian, Changshu Hospital Affiliated to Soochow University (First People's Hospital of Changshu), Changshu, 215500, China; QIAN Fu－liang, JIANG Zhi, ZHU Fei, SHEN Bai－rong, Center for Systems Biology, Soochow University, Suzhou 215006, China*

〔**Abstract**〕 Translational biomedical informatics is a new discipline integrating bioinformatics, medical informatics with translational medicine. The paper discusses the rise and ecology of the discipline, pointing out that full sharing of data is the foundation, complete genotype－clinical phenotype is the core and precise prediction that solves complicated diseases is the key, and analyzing the connotation of the discipline and development trend against the big data background profoundly.

〔**Keywords**〕 Bioinformatics; Medical informatics; Translational medicine; Biomedical big data

1 引言

随着基因组学、生物信息学和精准医学的迅猛发展，生物信息学、医学信息学与转化医学发生了交叉融合的趋势，多种分子组学的数据分析必须结合细胞、组织、器官、个体乃至群体层面的信息，才可能做到精准预测复杂疾病的发生发展，预测药物治疗、放化疗和免疫疗法的适用人群。在 2008

〔修回日期〕 2018－01－12

〔作者简介〕 白晋伟，硕士，馆员，发表论文 5 篇。

〔基金项目〕 国家自然科学基金资助项目“前列腺癌演变过程中的关键基因和模块及其作用机制”（项目编号：31670851）。

年左右，国际学术界提出了转化生物信息学的概念[1]，其目的就是将生物信息学发展起来的数据、模型和软件应用到转化医学的问题。从概念上讲生物信息学主要是人类基因组计划及多种分子组学技术发展所形成的一门新型学科，而传统的生物医学信息学在内涵上比生物信息学要广得多，学科所涉及的不只是分子层次，还有细胞、组织层次的图像数据分析、病人个体层次的临床电子病历分析、群体层次的流行病学数据分析等多个层次的数据建模、分析及其应用。因此从 2012 年起提出转化生物医学信息学的概念，召开系列的国际会议，出版专著和特刊来推动该学科的发展[2-4]。尤其是在当今大数据与精准医学的背景下，转化生物医学信息学的发展显得尤为重要。

在 PUBMED 数据库中分别以 Bioinformatics，Medical Informatics 和 Biomedical Informatics 为关键词查询在作者单位和题目、文摘中出现的次数，见图 1，其中右图为加 China 作为地址的限制词。

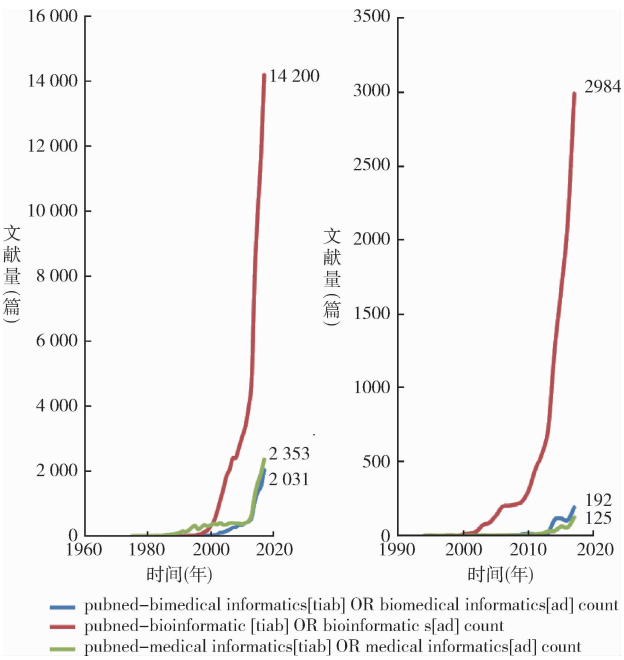

图 1 生物信息学、医学信息学和生物医学信息学发文趋势

由图 1 可见，过去 20 年来生物信息学领域发展十分迅速，根本原因是大量的基因组学数据及其后来的多种组学数据的迅猛增加和公开化。而生物医学信息学涉及的其他层次的数据如医学图像、电子病历等，由于隐私性和复杂性，公共数据库十分有限。医学信息学、生物医学信息学相对于生物信息学科的发展十分缓慢，但 2014 年起国际上医学信息学和生物医学信息学方面的文章陡然增加。而中国这方面的发展依然迟缓。随着个性化医学、P4 医学及精准医学的广泛推广，个性化临床表型数据将成为精准医学实现的瓶颈问题。大力发展转化生物医学信息学科成为必然。

2 转化生物医学信息学范式的兴起

2.1 近 20 年生物医学研究范式演变

科学研究范式的概念是美国著名科学哲学家托马斯·库恩 1962 年在《科学革命的结构》中提出并系统阐述的。它是指一定时期内在研究方法论基础上形成的研究原则和方法体系，库恩认为科学的进展是断断续续的，旧的范式和思考世界的方式是通过革命推翻的，而不是逐渐演变的。近 20 年来生物医学研究范式的演变情况，见表 1。

2.2 转化生物医学信息学范式

科学研究范式的演变有其必然规律，人类基因组计划实施推动了高通量技术的发展，导致在相对短的时间内发现大量的人类基因和生物功能元素；也促进转录组学、蛋白质组学、代谢组学等各种组学技术的发展，对各种组学数据的收集整理和模式发现推动生物信息学这门学科的产生。但是人们发现即使找到了生物系统中所有的成分，还是难以解释很多的复杂生命现象，因而产生现代分子系统生物学这门学科。由于人类基因组计划的目标之一是解决人类疾病的问题，生命科学的高速发展对医学的影响不如想象的那样大，因此人们进一步提出了转化医学的科学研究范式，强调结合临床病人标本和临床问题开展深度研究。转化生物医学信息学也随之产生，其目的在于将生物信息学发展产生的数据库和方法应用到转化医学研究中去促进临床的发现和应用。P4 医学与精准医学的产生是由于测序技术的普及，个人基因组及其他分子组学技术的实现

和成本降低，使得人们可以个性化预测、预防疾病。由于医疗数据和资源的缺乏、参与性医学也被提出并推广。这些学科范式都是在基因组技术发展的基础上演变而来，基因研究是其核心驱动力。然而疾病的产生是复杂多样的，是基因、环境、生活习惯多因素共同作用的结果，仅有基因和分子组学数据，往往难以全面地描述复杂疾病发生发展的全貌，临床表型组学数据，包括多种医学图像数据、电子病历数据、流行病学、生活习惯、环境因素等暴露组学数据则成为破解复杂疾病的下一个关键要素，因此将生物信息学、医学信息学融合起来的现代生物医学信息学学科应用到转化医学的研究，形成了一个新兴的科学研究范式：转化生物医学信息学。

表 1 近 20 年来生物医学研究范式的演变

范式	提出的问题	解决的手段	结果的解释
基因组学等各种分子组学	如何在系统中大批量发现功能元素	高通量测量技术	大量功能元件的发现
生物信息学	存在什么样的模式	数据库、计算模型	生物数据中模式发现与统计验证，基于数据库的解释
系统生物学	组分如何作用	系统论控制论方法	组分之间存在协同作用
转化医学	临床应用问题	结合临床标本验证	临床效果评价
转化生物信息学	数据如何在临床得到应用	临床数据 + 数据库 + 软件工具	临床疾病的分子基础及临床应用
P4 医学	如何做到个性化医学与疾病预防	整合多元信息建立系统模型	复杂疾病的异质性、进展、控制分析
精准医学	如何对不同病人用不同方案	个人基因组 + 各种的组学数据	疾病精准分类、慢性演化机制
转化生物医学信息学	如何建立精细的临床表型与基因型关系	众包，参与性医学、传感器信息、即时检验（point – of – care testing）等	理清基因型、表型、暴露组之间的复杂关系

3 转化生物医学信息学发展的学科生态

3.1 概述

交叉学科的发展需要一个好的学科生态环境为其提供必要的营养成分，转化生物医学信息学学科特征以及所需要的不同学科营养成分，见图 2。学科本身的关键是建立详细的基因型和表型关系网络，学科涉及的最相近的学科如生物信息学、医学信息学、转化医学本身也是交叉学科，因此这一学科所要求的知识背景和内容更加广泛、教育科研的挑战更大，需要从业者有较长时间的积累和较强的学习能力。

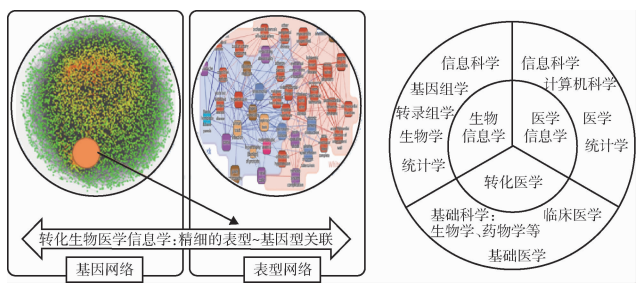

图 2 转化生物医学信息学以及学科发展生态

3.2 充分的数据共享是基础

过去 20 年生物信息学科得到迅猛发展，是因为有大量公共生物学数据库，尤其是各种组学数据库。当涉及医学时数据往往与个人隐私相关，个性

化的数据往往很容易被识别，从而暴露个人乃至家族的隐私性；另一方面，由于医学数据的来源广泛多样，标准化也十分重要。医学和健康信息数据的公开共享是一个难题，美国国家生物医学计算中心与医学密切相关的实验室就是关于数据隐私保护和共享的[5]。目前在数据隐私性及标准化方面已经开发了不少模型和算法[6-7]，只有数据共享的前提下，有大量的临床和病人相关的生物医学数据，转化生物医学信息学才有真正的发展基础。

3.3 完备的基因型－临床表型数据是核心

人类基因组计划促进基因数据采集的发展，然而对疾病临床表型的分类和描述依然很粗粒化，精细的临床症状、疾病信号的描述，与基因信息形成对应，才能形成完备的生物医学大数据而有效地应用于临床发现。过去由于疾病分类的粗粒化，很多重要的个性化疾病信息被统计平均掉了，只有基于足够大、具有精细关系的基因型－临床表现型数据，才可能构建精细的疾病谱，从而进行精准的诊断与治疗。完备的基因型与临床表型数据关系和结构是转化生物医学信息学学科发展的核心。

3.4 解决复杂疾病的精准预测问题是关键

科学问题是任何一个科学范式所要面对的第 1 要素。转化生物医学信息学在完备的基因型表型信息的基础上可以建立合理的模型、精准预测疾病的发生发展，甚至早期监控疾病的发生。由于疾病发展是一个复杂的动态变化，是基因与环境、生活习惯等相互作用的结果，利用生物医学大数据回答转化医学中的科学问题，需要信息学、生物医学和临床工作共同探求、深入交叉，脱离临床实践的闭门造车式的研究是难以提出复杂疾病精准预测、治疗和预后等方面的科学问题的。因此形成交叉学科的学术氛围和团队对学科发展十分必要。

4 转化生物医学信息学学科内涵

4.1 相近学科之间的关系

转化生物医学信息学与生物信息学、医学信息学等学科有密切的关系但又有其自身的特点，见表 2。这些学科的共同点都是利用信息学、计算机技术来收集、存储、分析和应用于相关领域，其不同点在于所研究的数据内涵是不一样的，转化生物医学信息学是研究基因型－临床表型二者共存的完备成对的数据，目前这样的数据是比较粗粒化和粗线条的，精细的结构关系依赖于临床医生和临床实践者，包括病人的参与才可能收集到。

表 2 相近学科异同

学科	内涵
生物信息学	利用计算机科学、数学、生物等知识分析和解释生物学数据
转化生物信息学	将生物信息学的方法及其技术应用于转化医学的研究
医学信息学	利用计算机、生物医学知识分析医学数据，主要是临床表型如医学图像、电子病历等
转化生物医学信息学	利用计算机、信息学技术，整合分子组学数据与医学表型数据寻找精细的组学－表型结构与特征，应用于临床诊断、治疗等

4.2 数据类型和数据特征

表 3 列出了一些生物医学数据中的常见数据种类，除此之外随着现代技术的不断发展将来还会涉及音频、视频、网页结构、信号波等其他类型数据类型。图 3 总结了生物医学数据的 5 大特征：（1）隐私性。即信息的个人隐私安全问题。（2）异质性。即对于同一疾病具有不同的数据特征和关系。（3）多层次。如分子层次、细胞层次、组织层次、个体层次、群体层次等。（4）演变性。由于生物系统具有动态演变性，从而生物医学各种数据之间又有一定的内在演变规律。（5）系统性。活的生物系统历来就是一个复杂系统，生物医学大数据将来可以帮助重构生物动态系统，而生物系统具有很强的鲁棒性特征。

表 3 生物医学数据类型及分析方法

数据种类	实例	分析方法
字符串数据	DNA/RNA 基因组学	字符串匹配（正则表达）、模式分析
图像数据	医学图像（EM，MRI，CT 等）	分割压缩、融合、可视化等
文本数据	电子病历等	自然语言处理等
数值数据	代谢物基因表达，蛋白质表达等	常用的统计计算等
相互关系网络数据	蛋白质、基因相互作用网络等	网络分析方法、中心度、度分布、鲁棒性表征等

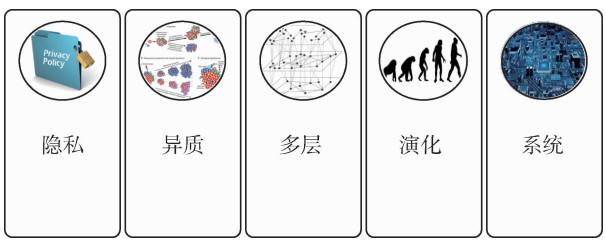

图 3 生物医学数据基本特征

4.3 交叉学科不是几个学科的简单相加而是融合

学科间的交叉从来不是简单的学科加和，将医学知识和信息知识加起来不是医学信息学，医学信息学必须是将医学知识与信息分析融合，产生医学信息特有的科学方法和科学问题才能成为一个真正的交叉学科。转化生物医学信息学不是简单的生物信息学及基因组数据分析与医学信息学的简单加和，而是将基因型与临床表型融合到一起去解决临床医学的精确诊断、预后、治疗等问题，构造学科特有的数据库、模型和软件工具，提出该学科特有的理论。

5 大数据背景下的转化生物医学信息学

5.1 数据与科学问题

数据科学家的研究对象是数据，有怎样的数据便可以用于提出什么样的科学问题，例如一个只包含癌症“有”或“无”的信息数据，可以利用此数据建立模型用于对未知数据进行分类预测，但因为没有关于药物使用效果的信息，这样的数据是难以用于回答药物疗效方面的科学问题的。过去通常在采取数据之前就有一些基本的科学问题设定，根据预先设定去尽量收集相关的信息。在大数据时代有时收集信息时对科学问题并不预先知晓，因此会尽量采集更广、更大量的信息，以便用于将来需要。随着 P4 医学中第 4 个 P 即参与性医学（Participatory Medicine）的兴起和众筹文化的流行，社会网络概念的应用，尤其是可穿戴传感器与互联网应用促进生理数据的实时动态收集，个性化医学数据的收集将越来越容易[8]，真正的生物医学大数据时代即将成为现实。

5.2 模型与关键数据

大数据的真正价值有时需要建立好的数学模型或者智能模型才能得到。生物医学数据常常是相互关联的，真正有用的关键数据可能只是少数。即从大数据中寻找到可以操作的小数据应用于临床是转化生物医学信息学的重要任务。如生物标志物、疾病的风险因素以及疾病的驱动变异寻找等。另一方面由于复杂疾病的发生发展有其客观的分子机理和规律，这些生物医学的基本规律在大数据分析和建模中可以帮助验证模型可靠性，同时也有助于提出合理的模型。生物医学中的基本规律包括遗传学、物理以及化学的基本原理等，也包括统计学中的数据分布规律等。

5.3 大数据时代的转化生物医学信息学

目前工业革命正在从信息技术时代向数据时代转变，大数据时代给转化生物医学信息学发展提供 3 个方面的基础：（1）大数据文化与思维。当全社会普遍认为下一代工业革命是由大数据所带来时，将自然接受生物医学大数据的范式并加以推广和应用。（2）为生物医学大数据的采集、存储、提取和分析提供技术保障，如 NoSQL、Hadoop，MAPreduce 等。（3）由大数据所带来的深度学习、人工智能技术为生物医学大数据的分析提供模型软件工具

基础，如 Caffa、Tensorflow 等人工智能工具的普及为转化生物医学信息学的模型构建、智能挖掘提供了工具和软件框架。总之，转化生物医学信息学一定会在大数据时代背景下充分吸收大数据技术的成果来推动本学科的发展壮大。

6 结语

转化生物医学信息学新范式将充分整合分子组学数据与临床医学数据，尤其是精细的临床表型信息，通过精准地建模，预测疾病发生发展以及个性化治疗的效果，会给医学的发展带来意想不到的效果。尽管如此仍需不断地面临多种挑战，如目前具有生物、医学、临床、信号处理等交叉学科背景的从业人员依然很缺乏；教育和培养这样的多学科交叉背景的学生将成为首要任务。交叉学科的生态发展往往不是很容易，正如 Leory Hood 教授在讨论系统生物学学科发展时所说：交叉学科发展的挑战首先就是社会学的挑战[9]。学科深入交叉、内涵加深是发展这一学科的前提。随着大数据技术的发展与应用，同时面临老年化社会的到来，从临床治疗到健康管理将成为一个趋势。转化生物医学信息学的应用未来对于预测疾病发生趋势将发挥重要作用。

参考文献

1 Butte AJ. Translational Bioinformatics: coming of age [J]. Journal of the American Medical Informatics Association, 2008, 15 (6): 709 - 714.

2 Chen J, Lin Y, Shen B. Informatics for Precision Medicine and Healthcare [J]. Advances in Experimental Medicine and Biology, 2017, (1005): 1 - 20.

3 Chen J, Qian F, Yan W, et al. Translational Biomedical Informatics in the Cloud: present and future [J]. BioMed Research International, 2013, (2013): 658925.

4 Zhao Z, Shen B, Lu X, et al. Translational Biomedical Informatics and Computational Systems Medicine [J]. BioMed Research International, 2013, (2013): 237465.

5 Ohno - Machado L, Bafna V, Boxwala AA, et al. iDASH: integrating data for analysis, anonymization, and sharing [J]. Journal of the American Medical Informatics Association, 2012, 19 (2): 196 - 201.

6 Ohno - Machado L. Informatics 2.0: implications of social media, mobile health, and patient - reported outcomes for healthcare and individual privacy [J]. Journal of the American Medical Informatics Association, 2012, 19 (5): 683.

7 Wang M, Ji Z, Wang S, et al. Mechanisms to Protect the Privacy of Families When Using the Transmission Disequilibrium Test in Genome - Wide Association Studies [J]. Bioinformatics (Oxford, England), 2017, 33 (23): 3716 - 3725.

8 Bai J, Shen L, Sun H, et al. Physiological Informatics: collection and analyses of data from wearable sensors and smartphone for healthcare [J]. Advances in Experimental Medicine and Biology, 2017, (1028): 17 - 37.

9 Hood L. Leroy Hood Expounds the Principles, Practice and Future of Systems Biology [J]. Drug Discovery Today, 2003, 8 (10): 436 - 438.