解码“生命天书”

■本报记者 潘锋

2003 年 4 月 14 日,经由 6 国科学家历时 13 年努力的人类基因组计划宣告完成。科学家们宣布绘制完成了人类基因组序列图,获得了人体基因遗传密码的“生命天书”。

其实在这之前,科学家们就已经意识到,拿到“生命天书”并不意味着掌握了生命的奥秘,读懂“天书”才是关键。

后基因组时代,生命活动的执行者、生命现象的直接体现者——蛋白质,成为研究焦点。进行蛋白质组研究,成为读懂“生命天书”的重要途径之一。

蛋白质组研究的重要特征,是更大规模生物学数据的产出。如何解读数据、挖掘数据背后的生物学意义成为研究的基础与前提。生物信息学,正是在这一过程中展现出了越来越重要的能力和价值。

朱云平,军事医学科学院放射与辐射医学研究所研究员、北京蛋白质组研究中心生物信息学研究室学术带头人,他领导我国蛋白质组生物信息学领域最重要的团队之一,从最开始就参与到了解码“生命天书”的这一伟大事业中,并发挥了重要的基础支撑作用。

1

1995 年,蛋白质组(Proteome)这个词语第一次出现。它指某个时刻,某个组织、器官或个体中所有蛋白质的总体,是一个整体的概念。

人的基因组是相对稳定的,但人体每个阶段甚至每个时刻的蛋白质组都不相同。“就像蝴蝶,由卵变虫成蛹再破茧成蝶,无论形态如何变化,它都是蝴蝶,这是基因组决定的。而它在不同阶段表现出了不同形态,则是因为它的蛋白质组发生了变化。”朱云平在接受本报记者采访时如是比喻。

正是因为这些完全不同的蛋白质表达,才造就了生物形态的千变万化,也决定了生物功能的多样性。

蛋白质组学的真正含义在于:它不是按照传统的方式孤立地研究单个蛋白质分子的功能,而是应用各种蛋白质组学技术,研究蛋白质整体在复杂的细胞环境中的表达和变化模式。蛋白质组学旨在列出全部蛋白质的细目,弄清每一个蛋白质的结构和功能及蛋白质群体内的相互作用,对比在疾病和健康状态下它们的表达水平的变化,寻找疾病发生发展的规律。

“蛋白质是生命活动的执行者,每一种生命运动形式,都是特定蛋白质组在不同时间和空间出现并发挥功能的结果。人的各种状态都与蛋白质组的变化和功能密切相关。比如某些蛋白质表达的高或低,就能够表征人体可能处于某种疾病状态。”

朱云平指出,对蛋白质组的研究,不仅能为研究人类生命活动规律提供物质基础,而且能为众多疾病机理的阐明及攻克提供理论根据和解决思路。科学家可以通过对正常个体与病理个体的蛋白质组进行比较分析,找到某些疾病特异性的蛋白质分子,进而将其确定为新药设计的分子靶点或疾病早期诊断的分子标志。

因此,开展蛋白质组学研究,探索生命奥秘是目标之一,更重要的是服务人类健康。

2

就像当初的人类基因组计划一样,人类蛋白质组计划召集了世界范围内感兴趣的科学家来参与,仍是欧美主导。所不同的是,中国不再是后来者、追随者,而是从一开始就参与其中,甚至在某些领域处于领导地位。

1998 年,我国第一个关于蛋白质组研究的自然科学基金重大项目启动,主要进行平台建设和相关技术的先导性研究。

2001 年,军事医学科学院联合中国科学院上海生命科学院、复旦大学等多家国内该领域最具实力的科研院所,成功申请“人类重大疾病的蛋白质组学研究”“973”项目,并于第二年启动。该项目以肝炎、肝癌等影响我国人群健康的重大疾病为对象,开展蛋白质组学研究,创建支撑技术平台,构建相关理论和技术体系。

2002 年,以“973”项目的合作为契机,中国人类蛋白质组组织成立。同样在这一年,我国科学家在第一次国际蛋白质组大会上提出了开展人类肝脏蛋白质组计划的建议。

朱云平(左一)和他的生物信息学团队

肝病在世界上波及范围甚广,我国也是一个肝病多发国,中国人群的肝病发病率和死亡率一直都很高。提请开展肝脏蛋白质组研究,是出于对世界肝病的发生情况和我国国情的综合考虑。

2002 年 12 月 15 日,国际人类蛋白质组计划(HPP)启动。经过各国科学家讨论同意,“国际人类肝脏蛋白质组计划(HLPP)”作为首批两项行动计划之一,由中国科学家贺福初院士领导。

这是第一个人体组织器官的蛋白质组计划,也是由我国倡导并领导的第一个生命科学领域重大国际科技合作计划。“我们自己一开始并没有意识到,后来听科技部这样评价才知道。以前参与人类基因组计划,我国加入得晚,只承担了整个计划的 1%,而在国际人类肝脏蛋白质组计划中,由中国倡议并主导,所作出的贡献占到了 30%以上。”

无论是“973”项目,还是 HLPP,有一个课题一直包含在其中——蛋白质组的生物信息学研究与系列数据库的建立。它的存在,旨在为蛋白质组学研究提供生物信息学技术支撑平台,大量运用数理的方法分析生物学数据,寻找生物学规律。

3

蛋白质组学是实验科学,它的发展极大地依赖于实验技术的发展。蛋白质组学的快速发展,为生物信息学相关领域的崛起提供了机遇,同时在很大程度上,生物信息学的发展为蛋白质组研究提供了重要的支撑和新的思路,二者相得益彰,蓬勃壮大。

蛋白质组学是在基因组学基础上发展起来的,很多研究方法都借鉴了“前辈”。然而,基因组是一维的、相对稳定的,而蛋白质组是多维的、动态的,这就意味着,蛋白质组学的数据更复杂,数据规模更大,数据牵涉的领域更多,它的复杂程度至少比基因组研究高出 3~4 个数量级。

面对大规模、高通量的复杂数据,找准核心环节重要问题开展研究,是提高效率、领先于人的关键。对这些数据进行有效管理和“瘦身”,就是生物信息学这一新兴交叉学科出现和存在的最初原因。

生物信息学伴随基因组计划的实施而兴起,是一门综合利用生物学、计算机科学和信息技术,揭示大量而复杂的生物数据所蕴涵的生物学奥秘的学科。

尽管成长迅速,但蛋白质组学仍是一个“小小少年”,基础还不牢固,各方面也不成熟。其中一个表现就是,因为没有统一的标准,它所产生的大量数据的质量是参差不齐的。哪些有用?哪些是金子?一般人甚至一般的实验室都很难辨别,也没有能力辨别。这时就轮到生物信息学团队发挥作用了。

“在海量数据中,像淘金一样,把真正有用的部分筛选出来。首先把所有数据收集起来,然后对它们进行有效处理和质量控制,因为数据中会有很多噪音,我们要想办法把噪音去掉,从中识别出真正有用的信息,进行注释和分析,形成数据库并展示,为研究者提供便利的工具。”朱云平介绍。

4

在我国的蛋白质组生物信息学领域,如果推选代表者,朱云平和他所带领的课题团队肯定是名列前茅的。

朱云平毕业于国防科学技术大学核物理专业,计算机是他最常用的工具。毕业后被保送到军事医学科学院读研究生并从事放射医学研究,研究内容更偏重物理学科方面。毕业后第二年,即执笔申请并获得了军事医学科学院首个数理学部的国家自然科学基金课题,此后几年作为主要贡献人获得了两项军队科技进步奖二等奖、1 项国家科技进步奖三等奖。

朱云平一直对生命科学非常感兴趣。他认为生命科学最大的魅力在于其未知性,关乎最复杂的系统,又跟人自身密切相关,而“对人类自身奥秘的探索是最吸引人的”。

朱云平能够从放射医学研究转到蛋白质组生物信息学,首先得益于军事医学科学院开放、活跃的研究氛围,其次离不开他本人的努力,最重要的是时机得当。

上世纪末本世纪初,基因组计划进入收官阶段,成果不断,正是生命科学如火如荼发展的时候,“21 世纪是生命科学的世纪”也是那个时候叫响的。

那个时候,朱云平时常在思考——

“生命科学发展很快,前景很好,但它仍是实验科学,不是理论科学。随着生命科学的发展,在快速涌现和成熟的技术推动下,大量数据产生,我们是否能像牛顿定律、元素周期律一样总结出生命科学特征?把生命科学的发展规律也建立成一个完整的理论体系?这种可能有多大?

“人类基因组计划的完成,将为这一理论体系的建立打下很好的基础;生命科学技术的发展,也为其理论的构建提供了实现的可能。而理论体系的构建,需要有良好数理背景的人来参与。”

……

朱云平思考的结果,是提交给研究所的一份关于发展生物信息学的建议书。他的建议与院、所领导的想法不谋而合,也非常符合军事医学科学院的学科发展规划。

于是,兴趣、努力与机遇促成了朱云平职业生涯的华丽转身。2000 年 9 月,他从辐射防护研究室调到了基因组学与蛋白质组学研究室,并专门成立了课题组,开始转向蛋白质组生物信息学研究。

从国家第一个蛋白质组科研项目的立项申请,到多个重大项目的实施完成,以及中国人类蛋白质组组织的成立,军事医学科学院始终参与其中并起到主导作用。而朱云平所领导的生物信息学实验室,作为大集体的一分子,也在一个个重大项目的历练中不断成长。

5

在蛋白质组研究时代,生物信息学的成果集中体现在蛋白质组数据库。这些数据库将为学术界提供便利,为解读生命、找寻生命发展规律提供基本的平台。

在十几年的发展中,依托多个项目和课题,朱云平团队在支撑蛋白质组学发展的过程中,实现了相关技术和平台的进步,产出了成系列的、高质量的数据库。

他们建立了蛋白质表达、修饰、定位、相互作用等系列数据库,库中包括人体重要组织器官——肝脏的成系列的、世界上规模最大、内容最丰富的高质量的蛋白质组数据集。

这个数据集包括了人胎肝蛋白质表达谱数据、中国正常成人肝组织及细胞器蛋白质表达谱数据、中国人肝组织磷酸化、糖基化修饰蛋白质数据、人肝细胞系乙酰化修饰蛋白质数据、中国人肝脏蛋白质相互作用数据、模式动物 C57 小鼠肝组织及细胞器蛋白质表达谱及肝组织、细胞器磷酸化蛋白质数据等。

同时,这些数据库通过一定方式实现了数据共享,部分内容已与国际数据库实现了数据交换。

这是目前最全面的人类器官蛋白质组数据库,从蛋白质组成、定位、相互作用、修饰等多方面提供了系统的信息及直接的实验证据,为全面解读人类基因组,揭示生命信息从基因组到蛋白质组的调控规律提供了重要基础。

除了以上用于实验数据管理的数据库,他们还建立了国际上最全面的肝脏综合知识库,收集整理了肝脏相关的基因、转录组、蛋白质表达、蛋白质相互作用、代谢、疾病关联等数据,能够为肝脏研究提供一站式服务。

2005 年,北京蛋白质组研究中心大楼落成,成为 HLPP 执行总部。朱云平团队的生物信息学实验室,也成为了 HLPP 的数据中心。

这就意味着,无论 HLPP 旗下的国际项目,还是 CNHLPP 旗下的国内项目,只要是关于肝脏蛋白质组研究的数据,都会汇集到这里,“由我们进行收集、整理、质控、整合、分析并建立数据库”。

6

2012 年底,历经 7 年酝酿和论证的“凤凰工程”启动。

国家蛋白质科学基础设施在北京和上海两地建设,“凤凰工程”是北京设施的代称,寓意科技工作者在生命科学领域不畏艰难、追求卓越,孜孜以求、不断创新。

这是一项军民结合、协同创新工程,旨在构建以蛋白质组分析系统、蛋白质结构解析系统、蛋白质功能研究系统为核心,以生物信息学、蛋白质大规模制备等分系统为支撑的蛋白质科学研究平台,建设蛋白质研究领域高端人才的培养平台,以及我国大规模蛋白质实物库、数据库和信息中心。

以 12 个亿的总资金投入到蛋白质科学研究的基础平台建设,足以体现国家和相关部门对这一领域的重视。

对于我国生物信息研究来说,“凤凰工程”将为他们带来许多优势——

“这将是国际蛋白质组领域最大的一个计算平台,硬件软件都是成系统、配套的。

“我们会把这些年自己研发的分析软件、计算方法等技术和工具都移植到这个大的平台上,系统完成后,它将是国际上最先进的蛋白质组生物信息学平台,没有之一。

“这会是国际上最大的一个蛋白质组数据平台,而且它的所有数据质量都是高标准的、可控的、直接可比的。”

这同时预示着朱云平他们所构建的数据库要扩容了。未来,它将会聚成为一个面向国际蛋白质组领域的公开数据库。

在这里可以以最快的速度,找到最详实有效的实验数据。不仅是最终结果,“还可以从实验结果追溯到实验的每一个环节、每一个步骤,直到最原始的数据。”这种追根溯源的全过程展示,为数据使用者提供了便利,他们可以通过检验过程任一环节的数据或原始数据,来判断数据是否可靠,找到支持研究结果的更多、更直接的证据。

很显然,只有所采纳的数据可靠,后期的数据分析、功能分析所得出的结论才会可信。朱云平建立的数据库所收集的数据是高质量、成系列的,这一点是他们的独特优势。

不同实验室所开展的研究侧重点不同,而数据库汇集了这些实验室的数据,有效管理之后,就可以形成一个比较全面的数据网,更有利于进一步工作的开展。

在这个数据库的支持下,生物信息学的存在与发展,将产生更多的实际意义和价值:

“比如,汇集了不同实验室关于肝病不同发展时期的研究数据,我们就可以对肝病的整个发展过程进行从头到尾的分析,从而获得更全面的结论。

“比如,汇集了肝、胃、肠等脏器的肿瘤疾病发展信息,就可以分析不同器官肿瘤的发展过程中,有哪些分子在同时起作用,或者不同脏器的肿瘤发展存在哪些不同。

“再比如,我们可以整合消化系统几个器官的相关数据,分析它们之间存在哪些特征和关联。”

总之,有了成系列、高质量的数据资源,就能够找到存在生理、病理关联的内容为研究对象,可以更有目的地去比较和分析,进行功能上的研究,找到更多具有生物学意义的知识,更具系统性的生理、病理规律。

7

采访中,朱云平也谈到了现阶段开展生物信息学研究所面临的问题,简而言之两方面——人才、设备。

生物信息学是一个交叉学科,涉及到生物、医学、数学、物理、化学、计算机等多个学科。从事这一领域工作,要求研究人员具备多个学科的知识积累,最理想的知识结构状态是既能熟练运用数据分析的技术和方法,又精通生物学基础理论。

朱云平说,这对人的要求非常高。要面对海量复杂数据不发憷,能够运用工具分析出数据中隐藏的规律,能注意到哪些地方发生了统计学意义上的变化;还要能看懂一条条曲线所代表的生物学意义,并能深入解读产生这些规律和变化的原因,及其可能产生的影响等更深层次的生物学问题。

长期以来,生命科学领域基础研究经费的使用在两个方面限制较多——大型设备和人员。这对于以消耗耗材为主的传统生命科学领域本无不妥,但对于新生的生物信息学来说就束手束脚。海量数据的处理分析,离不开大型的服务器和数据存储设备,同时也需要很多软件研发方面的人才,设备和人才所需经费的欠缺,严重影响相关研究工作的开展。

另外一个人才方面的问题体现在评价机制上。工具、平台、数据库是进行生物信息学研究与应用的最重要的内容,这些工作基础性、工程性较强,在生命科学研究领域很难发表影响因子高的论文。在以研究为主导的单位,近年对一个科研人员的评价往往以发表论文的影响因子为导向,这就导致没有人愿意去做生物信息学相关的工程性工作,即使愿意做也不易生存,致使对大科学项目的支持有一定困难。

尽管如此,他们并未放弃努力。“好在有院领导、所领导的支持,给我们配备了素质最好的人才。”有一个时期,在朱云平的课题组,包括所有工作人员和研究生,来自于北京大学、清华大学、国防科学技术大学等国内顶级院校的人员数目,比军事医学科学院内其他任何一个研究所的相同来源的人员数目都要多。

在这个课题组,每个人都在努力学习,一直在相互靠拢、融合,同时,合作成为另外一条取长补短的途径。内部合作、外部合作、国际合作,他们利用一切机会学习和进步。

伴随“凤凰工程”的建设,一支高素质的软件研发团队也将打造而成。这样,困扰生物信息学研究的设备和人才问题都将逐步解决。

8

朱云平对每个想进实验室的学生都说过一句话:没有强健的体魄,没有坚强的神经,不要进入这个实验室。

当他们毕业的时候,他告诉他们:除非你确实热爱这个事业,否则别留在我们实验室。

身处国际前沿领域,与全世界的优秀科学家竞争,所承受的压力可想而知,工作强度可想而知。在这一领域工作,要扛得住压力,耐得住寂寞,更要经得起多学科知识的洗礼,它需要研究人员有强烈的求知欲,有较好的学习能力,还要有极强的合作精神和奉献精神。

求学是为了成长,这里正适合年轻人的锻炼,但体力和精力都要足够强大;而毕业去向则关乎今后的生活,“去任何地方任何单位,过日子都会比我们这里舒服。”所以,才会出现这样截然不同的两句话。

正因为这样严苛的要求,才成就了这支生物信息学领域的领头团队。“很多东西都是我们从头做起,现在已经建立了蛋白质组研究生物信息分析的一系列算法、方法、工具、平台和数据库,我们有了成系列、成体系的东西。与此同时,我们一直在参与国际、国内大型的科研项目,并在其中起到了很好的支撑作用。”朱云平说,近几年,他们的工作也得到了国家和业界的更多关注和肯定。

2008 年,朱云平课题组参与的“蛋白质组支撑技术及其在人类重要疾病与生理过程研究中的应用”项目,获得了 2007 年度北京市科学技术奖一等奖。2012 年,参与的“蛋白质组技术与重要生理和病理过程的蛋白质组研究”获中华医学科技奖二等奖。

2012 年底,由生物信息学课题组为主并牵头完成的“蛋白质组信息学方法的研究及其支撑平台的构建和应用”项目,获得了中国电子学会电子信息科学技术奖一等奖。这是该奖项设立 10 年以来,唯一授予生物信息领域的一等奖。但从获奖到现在,课题组没有举行任何的庆祝活动,甚至实验室内部都没有为此而聚餐一次,一切都还是那么平静、自然。

现在,他们联合了太仓市生命信息研究所、重庆邮电大学和国家超级计算长沙中心等单位,正在建设一个宏大的蛋白质组数据库。这个叫做“iProX”的数据资源库,将接收全世界的蛋白质组研究数据,打破欧美在生命科学数据领域的垄断。其最基本的用途,是为研究者提供期刊论文数据的提交场所。相对于欧洲的 PRIDE、美国 PEPTIDE ATLAS,iProX 主要立足亚太地区并与他们进行数据交换,共同为国际学术界提供公共的数据资源,方便用户就近访问。4 月 18 日,在英国利物浦召开的国际蛋白质组数据联盟大会上,iProX 的报告获得了大会的充分肯定。数据联盟的负责人说:“一旦你们准备好,请马上告诉我们,然后采用国际统一的序列号,我们就一起工作了。”iProX 将是我国在国际上影响力最大的生命科学数据库。

近年,“生命组学”蓬勃发展,各个层面的组学都将带来海量数据的产出。如何对这些数据进行有效的分析整合,仍是现阶段生命科学研究的热点,也是难点。

我们有理由相信,随着信息技术的迅猛发展,生物信息学必将在解码“生命天书”的过程中起到越来越重要的作用。生物信息学大发展的时代,即将到来。

北京蛋白质组研究中心