

# 系统生物学研究进展\*

杨胜利

(上海生物工程研究中心 上海 200233)

**摘要** 人类基因组计划把生物学由分子生物学时代推向系统生物学时代,生物学由分解转向整合,由叙述科学转变为定量预测科学。

**关键词** 系统生物学,组学,计算生物学



20 世纪生物学经历了由表及里、由宏观到微观的发展历程,人们对生物体的认识也逐步由表型深入到分子层次。但生物体是一个复杂系统,只有通过各种分子机制、途径和网络的整合才能全面、系统地阐明复杂的生

物学现象。1990 年启动的人类基因组计划是生命科学史上第一个大科学工程,该计划是生物学发展的一个重要转折点,使研究工作由分解转向了整合,研究的构架也由单一的生物学实验室转变为大科学工程与传统生物学实验室相结合的模式。生物学与数学、物理、计算机科学将更紧密地交叉,使生物学由描述性科学发展为定量预测的科学。在人类基因组计划带动下出现的一系列组学(Omics),逐步把分子生物学时代推向系统生物学时代。

系统生物学是在细胞、组织、器官和生物体整体水平上研究结构和功能各异的生物分子及其相互作用,并通过计算生物学来定量阐明和预测生物功能、表型和行为。系统生物学将在基因组测序基础上完成由 DNA 序列到生命的过程,这是一个逐步整合、优化的过程。系统生物学的发展预计需要一个世纪或更长时期,因此常把系统生物学称为 21 世纪的生物学。

## 1 系统生物学的组学技术平台

随着基因组测序的进展,科学家逐步将目光集中到功能基因组研究。在基因组水平研究基因的表达及其调控、基因及其产物的结构与功能、分子的相互作用、基因型与表型的相关性,推动了一系列基因组水平的测试技术平台,主要有基因组学、转录组学、蛋白质组学、代谢组学、相互作用组学和表型组学等,这些高通量的组学实验平台构成了系统生物学的大科学工程。

### 1.1 基因组学

基因组学包括基因组测序、基于单核苷酸多态性的基因分型、表观基因组学等技术平台。

基因组测序揭示了每一个生物体中所有的遗传密码,是系统生物学的基础和起点,因为基因组的序列决定了每一个生物体的基本性状,并与环境因子相互作用决定生物体的行为。2003 年已完成了人类基因组计划中人和 5 种模式生物的全基因组测序,同时也完成了多种重要动、植物品种和 100 多种病原微生物和工业微生物基因组的测序。

基因分型的精度已达单个核苷酸水平,2002 年由美国国立卫生研究院牵头启动了单体型组学研究,系统测定不同人种的单核苷酸多态性(SNP),这是一个比人类基因组计划更宏大的计划,两个无亲缘关系的个体间平均每 1 000 个核苷酸中可能有 1 个差异,这样任何个体之间可能有 300 万个 SNP。2001 年用于 SNP 研究的费用高达 1.58 亿美元,预计到 2005 年将达 12 亿美元。SNP 的研究将有助于

\* 收稿日期:2003 年 12 月 26 日



阐明疾病易感性和药物应答及其它性状、行为的个体差异,是实现预测医学和个性化医学的关键。

表观基因组学研究染色体上核苷酸修饰与基因表达及基因功能的关系,目前主要集中于 DNA 甲基化的研究,DNA 甲基化参与基因表达和功能调控,与细胞的分化和发育、肿瘤的发生和发展有关。

## 1.2 转录组学

转录组学是在 mRNA 水平研究基因表达谱,为用 DNA 芯片测定 mRNA 的丰度提供了一个可在基因组水平上进行高通量、平行检测的工具。高密度的 DNA 芯片可在 1 平方厘米上放置 2.5 万个寡核苷酸探针或 1 万个 cDNA 探针进行 mRNA 丰度测试。Affymetrix 的基因芯片 Human Genome U133 (一组两片)有 100 万个寡核苷酸探针,代表 3.3 万个人基因。Motorola Life Sciences 已推出人、大鼠、小鼠的基因芯片,每个芯片有 1 万个全长 cDNA 探针。已用 DNA 芯片测定了不同细胞、组织的基因表达图谱,比较疾病和正常组织的表达图谱,细胞在不同分化、发育阶段和不同环境条件的基因表达图谱,这些结果不仅用于新基因功能、信号传导系统与细胞和生物体对环境因子的应答等研究,而且还可用于新药靶基因的鉴别和验证、疾病的诊断和预警、毒理和病理的研究。

## 1.3 蛋白质组学

蛋白质组学研究细胞和生物体的蛋白质分子的结构、时空分布和功能,可分为表达蛋白质组学、功能蛋白质组学和结构蛋白质组学三大分支。

表达蛋白质组学是蛋白质水平的基因表达谱,最常用的是双向电泳-质谱系统,用双向电泳、蛋白质染色、质谱分别作为分离、检测和鉴别方法(MS)。目前这个系统用于低丰度蛋白质和膜蛋白的分析仍有一定困难,一般仅能检测 2 000—3 000 个蛋白质。近年来发展的二维色谱-质谱、蛋白质芯片-质谱系统进一步扩展和提高了蛋白质组研究的能力。表达蛋白质组已用于特定生理和病理状态下蛋白质标记分子的发现和鉴别,这些蛋白质标记分子可用于药物发现、临床诊断和预警及疾病的监控。

功能蛋白质组学研究难度较大,涉及蛋白质的

翻译后加工、转运和亚细胞定位、蛋白质的构象和相互作用等。已发展用基于蛋白质活性免疫探针和化学探针方法检测和富集活性蛋白质,如用免疫探针或化学探针富集磷酸化蛋白、用化学探针标记蛋白质组中 6 类机制各异的酶超家族、用化学探针监控蛋白质组中酶的活性、用酶活性中心专一的化学探针筛选酶的抑制剂等。用绿色荧光蛋白标记的蛋白质组研究蛋白质的亚细胞定位也已广泛应用。

结构蛋白质组学,亦称结构基因组学,主要研究高通量测定蛋白质的高级结构。除了高通量的基因表达和蛋白质纯化技术平台外,高通量的 X-射线蛋白质晶体结构测定和高通量的核磁共振(NMR)蛋白质溶液结构测定是结构蛋白质组学的关键技术平台,尽管目前通量还不小,但已取得较大进展。1999 年报道的蛋白质立体结构约 1 500 个,至 2001 年已达 1.3 万个以上。利用蛋白质结构知识已成功设计和开发了基于蛋白质结构的抗流感病毒药物和抗艾滋病病毒药物。膜蛋白的立体结构是药物学家十分关注的,因为膜蛋白约占人蛋白质的 30%,而药靶分子中有一半以上是膜蛋白包括受体和离子通道,但至今报道的膜蛋白立体结构为数甚少。

对模式生物啤酒酵母已进行了系统的蛋白质组研究。6 243 个酵母开放阅读框架(ORF)分别用串联亲和标签(TAP)和绿色荧光蛋白标签进行标记,使酵母蛋白质组的蛋白可在天然环境表达和定量分析。结果 6 109 个 TAP 标记的蛋白中有 4 251 个(约占 70%)在对数生长期得到表达,6 234 个绿色荧光蛋白标记的酵母蛋白组中 4 156 蛋白(约 75%)进行了亚细胞定位分析,将这些蛋白定位于 22 个不同亚细胞区域。

蛋白质组学研究有其特殊重要性,因为生物体的功能、表型和行为主要是由蛋白质及其相互作用决定和执行的。由于蛋白质结构和功能的多样性,蛋白质组学不可能像基因测序和转录组学有一个通用技术平台,目前蛋白质组学的技术平台还远不能满足研究需要,这一方面将会有较大发展。

## 1.4 代谢组学

代谢组学的英文名有两个即:metabolomics 和



metabonomics。前者研究单个细胞或单一类型细胞的代谢调控和代谢流,常用于植物和微生物。后者通过分析体液或组织测定整个机体的系统代谢图谱和功能调控,一般用于人和动物。

代谢图谱的测定可用核磁共振、质谱、气相色谱(GC)、液相色谱(LC)或这些方法的组合,如 GC-MS, LC-MS, LC-MS-NMR 等。NMR 和 MS 是产生多变量代谢数据的有效方法,可提供较多的分子结构信息。MS 灵敏度高,而 NMR 是一种无损伤测试方法,可用于复杂混合物和完整组织及细胞的生化分析。荧光共振能量转移分析法(FRET)可实时测定代谢物的变化和亚细胞定位。

代谢处于生命活动调控的末端,因此代谢组学比基因组学、蛋白质组学更接近表型。基因组学、蛋白质组学研究一般是发现和鉴别潜在的可能性,而代谢组学研究则是发现和鉴别真实的变化。代谢组指纹图谱已成功地用于冠心病和肿瘤的预警与诊断以及药物毒性和疗效的评估。代谢组学的研究还可用于代谢相关基因沉默突变的发现。植物和微生物的代谢组学研究已在指导代谢工程设计和提高目标代谢产物的产量等方面取得了良好进展。

### 1.5 相互作用组学

生命活动中的各种功能和行为不是由单一分子完成的,都是通过体内各种分子的相互作用实现的。现在研究较多的是蛋白质间的相互作用。相互作用组学系统地研究各种分子相互作用,包括蛋白质-蛋白质、蛋白质-核酸、蛋白质-代谢物的相互作用和这些作用形成的分子机器、途径和网络。

研究分子相互作用的技术平台有三类:(1)分子生物学平台,主要是酵母双杂交和噬菌体展示系统;(2)生物化学技术平台,主要有免疫共沉淀、交链和结合试验;(3)蛋白质芯片技术平台。用酵母双杂交法分析了啤酒酵母系统 2 039 个蛋白质之间的相互作用,结果鉴别了一个由 1 548 个蛋白质参与包括 2 358 个相互作用的巨型网络和几个较小的网络。用蛋白质芯片鉴别了啤酒酵母系统中 1.1 万个蛋白质-蛋白质相互作用。

相互作用组学研究可用于构建生物系统中的

各种途径和网络,鉴别参与网络和途径的生物元件,形成系统生物学研究中的模块,进一步通过模块的相互作用研究构建完整生命活动线路图。

### 1.6 表型组学

目前,表型组学主要是细胞水平的研究,因为细胞作为生命活动的基本单元具有活生物体的主要性状,如信息传导、时空组织、繁殖、体内平衡和对环境变化的应答和适应等,而且可用细胞进行高通量全基因组水平的研究。

表型组学研究的主要技术平台是细胞芯片和组织芯片。细胞芯片是在全基因组水平对每个基因进行各种基因操作,包括基因敲除、基因导入、基因抑制和基因激活,构建相应的细胞株,并植入细胞芯片进行高通量的表型研究。组织芯片主要用于高通量的药理、毒理、病理研究。

表型组学是系统生物学组学平台的终端,通过基因组学、转录组学、蛋白质组学、代谢组学、相互作用组学到表型组学完成了由基因组序列到基本生命活动的全过程。已对大肠杆菌和酵母的糖代谢进行了从基因组到表型组的系统研究。细胞芯片也在新药和药靶的发现、新药评价等方面得到愈来愈多的应用,使新药研发从高通量过程逐步演变为高内涵(high content)过程。

## 2 计算生物学

阐明和定量预测复杂的生物系统需要实验研究和计算研究的整合。计算生物学通过建模和理论探索,为提出科学问题提供了强有力的基础。

生物系统是一个复杂系统,但与以往的复杂系统又有本质差别。其它复杂系统是由大量的简单元件相互作用而产生复杂的功能和行为,而生物系统则是由大量的功能各异且常常又是多功能的元件在选择性和非线性的相互作用下产生的复杂且是相干的功能和行为,对计算科学是一个新的挑战。

计算生物学可分为知识发现和基于模拟分析两个分支。知识发现也称为数据开采,是从系统生物学各实验平台产生的大量数据和信息中抽取隐含其内的规律并形成假说。基于模拟的分析是用计



算机验证所形成的假说,并对拟进行的体内、体外生物学实验进行预测。知识发现已在生物信息学中广泛应用,如根据核酸序列预测内含子、外显子、所编码蛋白质的立体结构;由基因表达谱推导基因调控网络等。这些研究是基于启发式的、统计学差异的分析方法。而模拟分析是致力于系统动力学的预测,大量高通量平台的定量数据的产生和积累为基于模拟的研究提供了坚实的基础。计算机建模和分析已提供了有用的生物学提示和预测,如细胞周期的分叉分析、代谢分析、生物振荡回路稳态的比较研究等。药物研发也是计算生物学的一个应用热点,虚拟筛选和药物分子设计,药物吸收、分布、代谢、排泄和毒性(ADMET)的计算机建模已成为新药研发的重要工具。

### 3 概念生物学

随着数据和实验结果的大量积累,用概念驱动方式分析这些结果并整合到可测试的途径和网络中而产生新知识和规律变得愈来愈重要。如不用概念驱动方法分析、组织这些数据,数据的大量产生和复杂性将成为研究的障碍。将各自独立的数据和事实综合为新的概念,就像把 26 个字母组合成词一样,这些较简单的概念可进一步综合成复杂概念,正如由词组成句子、章节和文章一样,这个过程就是概念研究。它与自动化的数据开采和传统的理论生物学不同,概念生物学对现有各种不同来源的信息和模型进行逻辑分析,产生一个假说,并将假说中的预测以可测试的方式表达,然后在公开发表的报道中寻找相关的信息。一般所研究领域的信息

愈多,进行概念研究的机会就愈多,成功率也愈高。

概念生物学将是高通量技术平台、计算生物学和传统生物学实验室的桥梁。

### 4 展望

世纪之交生物学逐步由分子生物学转向了系统生物学,由基因组序列到生命研究的进程比预期要快得多,人们真正全面了解生命的愿望正在逐步成为现实。系统生物学不仅是生命科学理论的重大发展,而且有极其广阔的应用前景,如应用于药物的发现和开发,有望改善新药研发的投入产出,疾病的诊断、预警和药物应答并逐步实现预测医学、预防医学和个性化医学;通过对农作物系统的设计和重构优化农作物品种;提升传统生物技术工业并开拓能源生物技术、材料生物技术和环境生物技术等新领域。系统生物学的研究将不仅对生命科学和生物技术产生重大影响,而且对整个国民经济、社会和人类本身产生重大影响。

#### 主要参考文献

- 1 Collins F S. A vision for the future of genomics research. *Nature*, 2003, 422:835-847.
- 2 Kitano H. Computational systems biology. *Nature*, 2002, 420:206-210.
- 3 Lockhart D J et al. Genomics, gene expression and DNA arrays. *Nature*, 2000, 405:827-836.
- 4 Palsson B. The challenges of in silico biology. *Nature Biotechnology*, 2000, 18(11):1 147-1 150.
- 5 Sumner L W et al. Plant metabolomics. *Phytochemistry*, 2003, 62:817-836.

## The Progress of Systems Biology

Shengli Yang

(Shanghai Research Center of Biotechnology, CAS, Shanghai 200233)

Human Genome Project is a turn of biology from molecular biology era to systems biology era. Biology is changing from reductionistic approach to integrative approach, from descriptive science to predictive science.

**Keywords** systems biology, omics, Computational biology

**杨胜利** 上海生物工程研究中心研究员,中国工程院院士。1941年1月出生于上海。1962年毕业于华东化工学院有机化工系。现从事基因工程、代谢工程、分子药理学和功能基因组学研究。曾获中国科学院科技进步奖一等奖、二等奖,上海市科技进步奖二等奖。