DOI:10.16262/j.cnki.1000-8217.2005.05.006

·学科进展与展望·

系统生物学的现状与展望

石铁流[1,2] 李亦学[1,3]

(1 中国科学院上海生命科学研究院生物信息学中心,上海 200032;
2 上海大学生物信息学中心,上海 200444; 3 上海生物信息技术研究中心,上海 200235)

[摘 要] 作为生物学的一个新领域,系统生物学的兴起及发展为研究生命活动提供了一个全新的思路和方法,它将各种组学的方法综合起来,通过高通量的实验手段,从转录组、蛋白质组、代谢组等方面多层次、全方位地对生物体内的生命活动进行检测,同时利用生物信息学的方法和手段对相关的数据进行分析及整合,并在此基础上,对相关的生命活动进行模拟,以期系统地研究和阐明生命活动的规律。本文简单介绍了系统生物学的现状,研究的内容及方法,并对今后的发展提出了一些想法。

[关键词] 系统生物学,生物系统建模,高通量,基因调控网络,蛋白质相互作用网络

20 世纪的生物学研究采取的是还原式方法(reductionistic approach)。现在看来,这种传统的单一、零散、小规模的还原论生物学尽管已发现了许多生命活动的规律,推动了我们对生命活动的认识,却难以有效地,系统地研究复杂生物体的生命活动。随着"后基因组"时代的来临,海量的生物数据不断产生,以及生物芯片、质谱仪等高通量技术的日渐成熟,使在收集、整合、数据挖掘的基础上全方位地研究生命活动的规律成为可能。以生物信息学和计算生物学引导的、以整体和相互关系为研究对象的系统生物学(Systems Biology)为此应运而生,成为当今生物学研究领域中的新的热点。

系统生物学这一概念的提出,为针对整个细胞和生物体内各种生理途径的研究指出了一个新的方向。与传统生物学方法孤立地研究某一基因或其编码的蛋白质不同的是,系统生物学是从整合的角度来研究生物学。整合的观点主要是认为生物体内各个子系统间是不独立的,从而希望寻找新的方法来阐明子系统间交互作用的问题。系统生物学和传统生物学的本质区别在于:它强调用系统的观点研究生物,即整体大于部分之和,更注重整体,注重各部分间的相互关系。系统生物学从不同的层次同时研究多重生物学信息之间复杂的相互作用,包括基因组 DNA, mRNA, 蛋白质, 代谢、信号传导途径, 基因调控网络和蛋白质相互作用的网络等,以期在此基础上理解它们之间是如何协同作用的,从而更清晰地阐明疾病的发生机理及药物的作用机制。因此,如果说上世纪的生物学研究采取的是还原式方法的话,那么,当今兴起的系统生物学采取的则是整合方法(integrative approach),是从叙述性的科学转向于定量、预测的科学。

1 系统生物学的研究目标

系统生物学的研究将从以下不同的层次来理解生命现象[1]:

(1) 理解系统的结构:如基因调控及生化网络,以及实体构造;

(2) 理解系统的行为:定性、定量地分析系统动力学,并具备创建理论或模型的能力,可用来进行预测;

(3) 理解如何控制系统:研究系统控制细胞状态的机制;

(4) 理解如何设计系统:根据明确了的理论,设计、改进和重建生物系统。

2 系统生物学的研究内容

目前系统生物学探索的,也是当今生命科学研

本文于 2005 年 5 月 23 日收到.

究的前沿，主要包括如下内容：细胞信号传导系统(signal transduction modeling)、基因调控网络(genetic regulatory network)、代谢途径(metabolic pathway)、蛋白质相互作用(protein-protein interaction)、生物分子标记的发现(biomarker discovery)、药物的筛选和药物效果的建模。

2.1 信号传导途径

信号传导过程控制着细胞的生存与凋亡。系统生物学方法期望通过建立细胞信号传导过程的模型，从而找出参与此过程的各个蛋白质间的相互作用的网络，阐明其在基因调控、疾病发生(如肿瘤等)中发挥的作用，为治疗这些疾病的药物发现提供依据。近几年来，对信号传导通路的定量分析逐日升温。如何建立数学模型来完成定量分析逐渐成为焦点，用连续性系统动力学方法来解决这一问题已成为这一领域普遍研究的方法。迄今，人们对以下几种信号传导系统的机制作了研究：磷酸肌醇-钙系统[2]、有丝活化蛋白激酶系统(mitogen activated protein kinase system，简称 MAPK)[3]、JAK-STAT活化系统[4]、离子通道功能受体系统[5]。

2.2 基因调控网络

基因调控网络研究就是利用生物芯片等高通量技术所产生的大量基因表达谱数据以及蛋白质-DNA之间相互作用等信息，利用生物信息学及计算生物学的手段和方法，结合其他的实验结果来构建全基因组范围的基因调控模型。

所有转录因子和其靶基因之间的相互作用可以用一个有向图(directed graph)来表示。图的节点(nodes)表示转录因子和靶基因，图的边(edges)表示它们之间的调控关系。最终基因调控网络将是一个很复杂的多层次的系统。研究基因调控网络的方法目前主要有聚类，基于表达谱和 ChIP-chip 的分析方法，基于表达谱和启动子序列的分析方法以及机器学习(machine learning)的方法等几大类。

聚类分析[6]建立在一个假设基础上：具有相似表达谱的基因可能被相同的转录调控机制所控制(可能被相同的转录因子调控，具有相同的转录因子结合位点等)，可能具有相似的生物学功能。上述这些传统的聚类方法也存在缺陷：它们考虑的是基因在所有条件下的相似性，然而许多基因只能在一定环境下而不是所有的条件下都能够共表达。因此，为解决这些局限，一系列新的算法运用到聚类分析中[7]，这些算法对基因表达矩阵的行和列同时进行聚类，因此称为双聚类(biclustering)。

ChIP-chip 方法[8]是用抗某一转录因子的抗体做免疫共沉淀，并与基因芯片相结合从而得到可以和该转录因子相互作用的 DNA 片段。ChIP-chip 数据虽然能够对转录因子和启动子之间的相互作用提供直接的证据，但并不能证明结合位点在生理条件下的作用，也不能说明转录因子结合在 DNA 上所发挥的作用是正调控，负调控，还是根本不起作用。基于表达谱和启动子序列的分析方法[9]是通过对基序(motif)的协同性下一个统计学的定义，进而来识别有功能的基序组合，如果启动子序列包含一对基序的基因的表达一致性值比仅包含某一个基序的基因的表达一致性值高，就认为这一对基序是有协同作用的。机器学习方法包括线性模型(linear model)[10]、贝叶斯网络模型(Bayesian model)[11]、布尔网络模型(Boolean network)[12]、SVM(support vector machine)方法[13]等。线性模型假定网络中某一点的表达水平是依赖于其邻近各点表达水平的线性组合。贝叶斯网络模型基于贝叶斯条件概率理论，将相关的各基因表示为一张有向无环图的节点，因此更适合于研究局部的基因调控网络。布尔网络则将组成网络的各个基因节点用“开”或“关”两种状态表示。不论是信号传导还是基因调控网络，都可以采用布尔网络模型，每一个信号蛋白或每一个基因或者处于活化(active)状态，或者处于失活(inactive)状态。概率布尔网络(probabilistic Boolean network)是对布尔网络的改进，既保留了布尔网络的优点，又可以更加有力地处理不确定性。SVM 最近被引入到基因调控网络的研究中，经过相关的训练，SVM 给每一个由转录因子和目标基因组成的基因对确定一个概率，在给定合适的阈值的情况下，用这些概率来构建调控网络。

2.3 代谢途径

主要是从定性、定量两方面用多种方法分析代谢网络。根据建模的目的一般可以分为三种：稳态分析、动态分析、灵敏度分析。根据这些目的，产生了很多建模方法，并被运用于特定的代谢网络，得到了符合生物体内真实代谢的结果。流平衡分析[14]是分析稳定状态的典型方法，其数学本质是线性规划，根据生物体的具体情况确定目标函数，将反应速率的变化范围作为约束条件，解这个规划问题可以得到稳态时所有反应的速率(流)。代谢控制分析[15]用于分析参数的灵敏度，以偏导数定义灵敏度，缺点是只适用于参数小范围变动的情况。基本流模式[16]是能够达到稳定状态的最小的一组酶，所

有稳态流均可以表示成基本流模式的线性组合，所以它可以看成是稳态流空间的一组基(a set of basis)。得到了基本流模式，就可以刻画稳态流的所有特征。前面提到的方法都不需要用到代谢网络的动力学特征，这是它们的优点，使方法简单易行。然而也由于没有考虑动力学数据，获得的分析结果是有限的。动态分析，即变量随时间的变化情况，只能通过动力学模型才能获得。微分方程[17]作为传统的动力学模型，在代谢网络动态分析中有着相当重要的作用。建立微分方程的困难在于建立反应速率表达式的困难。同时，如要得到有实际意义的连续模型，还需要对各参数有准确地估计。这一点往往由于缺乏数据而难以实现。总之，不同的建模方法从不同的角度反映了代谢网络的特点和本质，使人们更好的认识了细胞代谢，但它们也都有各自的缺点和局限性。

2.4 蛋白质-蛋白质相互作用

传统的生物学实验蛋白质相互作用的方法包括酵母双杂交(yeast two-hybrid)、与质谱连用的亲和分离等方法。但这些实验室研究方法的一个缺点是对实验室的条件要求高，周期长，花费大，而且准确性不高，会出现大量的假阳性和假阴性结果，如已有的结果表明，有时酵母双杂交的假阳性超过 50%。在这种情况下，人们发展了许多生物信息学的方法[18]，一方面期望通过对数据的分析，从现有的实验数据中去掉假性结果，抽提出真正的蛋白质相互作用的信息；另一方面，利用计算的方法来预测、分析蛋白质之间的相互作用，以期对实验予以指导，最终将两方面的信息结合起来构建生命活动过程中的蛋白质-蛋白质相互作用网络。在此基础上，对未知功能的蛋白质进行功能研究，探索已知功能的蛋白质的新的功能；同时，为新的药物靶点的发现提供线索。

2.5 生物分子标记的发现

对各种疾病状态下的生物分子标记的研究一直是生物医学的一个重点，近两年成为了生物医学研究的热点，它与药物的发现及临床实验有着紧密的联系。差异表达的蛋白质可以作为生物体状态表型改变的指标或标记，因此，蛋白质标记分子可以用来检测疾病状态的变化，跟踪疾病的严重程度以及监测对药物治疗的反应。蛋白质生物分子标记的应用将有利于药物发现，临床前的毒理学研究，基础和临床研究以及疾病的诊断。疾病是一个过程，而不是一种状态，因此，单一生物分子标记可能很难作为检验一种疾病发生、发展的指标。目前大部分的相关研究人员认为采用多个生物分子标记才可能有效地检测疾病。寻找多个生物分子标记是目前这个领域主要的方向。

2.6 药物发现

这是系统生物学最大的应用热点之一。ADMET(absorption, distribution, metabolism, excretion, toxicity)是困扰国际制药界的主要难题。一个新药由于 ADMET 而招致的失败，其经济损失平均为 2 至 5 亿美元。问题的本质是：人们没有在生物系统层面上阐明所研究的先导分子与其在人体中靶点之间的相互作用。在临床前各研究阶段，由于代价昂贵，现有的新药研究是在独立于遗传网络和蛋白质的相互作用途径的情况下将配体小分子与生物大分子的作用优化，从而未能考虑到各个分子之间相互作用的协同性。现有的 ADMET 预测方法基于传统的结构与活性定量关系(QSAR)模型，而不是预测 ADMET 的分子生物学机制，因此效果不佳。ADMET 的分子生物学机制的研究及阐述对创新药物研究有着极为重大的意义。而系统生物学的研究方法为 ADMET 预测和新药的药理机制的模拟开辟了一个全新的途径。

3 系统生物学的研究方法

当今生物学的研究可以依据采用的途径分为两大类：一种是根据所观察到的生命活动现象提出假设，建立相关的模型，然后进行模拟，并设计实验来验证这个模型是否正确，这种途径可称之为 Top-down；另一种方式是经过大量的实验，积累相关的原始数据，再对数据进行分析、比较，此基础上，建立生物学模型，并对生物学模型进行模拟，再以实验来验证模型，这种途径称之为 bottom-up。可以看出，这两种方法针对生物体系统的研究起着殊途同归的作用。

系统生物学主要采用的是后一种研究方式，其研究对象是生物体内的网络结构，而不是简单的组成这个网络的单个分子组成成份。一个系统可以是一个基因调控网络，一个代谢途径，也可以是一个细胞、一个组织，甚至可以是一个完整的复杂的生物体。因为系统生物学需要同时对相互作用的各个组成成份进行研究，故高通量和定量的技术的采用非常重要。同时各种计算方法也是必须的，用来处理、分析了解复杂生物系统所必须的海量的实验数据。

系统生物学在进行实验设计时，研究的对象及

确定的测量的生命活动的时空点或许与过去比差不多，但是使用高通量的分析技术收集大量数据，同时，采用人工或相关的计算算法对现已发表的文章进行文本挖掘，以期从中找出各种基因和蛋白质之间的相关性及在不同实验条件下各种基因表达或蛋白质活性的变化情况，PubMed 已是文本挖掘的主要资源。而且分析数据的方法跟传统的生物学有时有着很大的差异，甚至会引入许多生物学者不熟悉的分析方式，包括各种复杂的数理统计方法、概率、算法等，以及基于这些数据上的数学建模。这种先收集大量数据，再做理论分析的研究方法应是更有效率的方式，可以在同一时间内从多种不同的角度来观察生物体内的各种活动，更重要的是可以有机会观察到同一时间内许多路径间的交互作用，这是传统生物学研究中无法实现的，从而为多方位的研究生物体内的生命活动提供了可能，有机会由整合的角度分析生物学。

4 系统生物学的发展趋势

目前，系统生物学主要是集中在对人的疾病的研究及新的药物的开发上，并取得了一些显著的成绩。系统生物学研究在微生物领域也在逐步展开，主要集中在对微生物内代谢途径的研究，通过对代谢途径的分析、比较，以期找出最佳的代谢调控方式，关键因子，反应的最佳状态，从而提高有效成份的产率。系统生物学在植物中的应用似乎要滞后于其在动物及微生物中的进展。笔者认为植物中叶绿体应是植物系统生物学的最好研究对象之一，针对叶绿体的系统生物学研究有着重要的经济意义及学术价值。我国在光合作用的研究领域已取得了世界瞩目的成绩，在传统生物学方面有着坚实的基础，我们应尽快地将系统生物学研究的思维及方法应用到叶绿体，从整体上研究光合作用过程中各种因子之间的相互作用，找出与光合作用相关的重要调控因素，以期对光合作用途径进行可能的优化，对于农作物的育种，提高其产量有着重要的指导意义；其次，植物的次生代谢途径也将是系统生物学很好的研究对象，有着重要的经济价值，因为中药的很多有效成份就是植物体内的次生代谢产物。同时，在研究中药的治病机理上，我们也需要引入系统生物学的研究思路，因为中药的有效成份是多分子，其在人体中的作用靶点也应是多样性的，各个分子对各个靶点的作用一定会存在协同效应。因此，只有全方位地从系统的角度出发，才有可能真正阐明中药的作用机理。

5 结 语

系统生物学已越来越受到生物学界的重视，2002 年 *Science*，*Nature* 分别发表了专刊或特别专题就此领域的进展和未来的发展方向展开了讨论，有关系统生物学的国际年会也吸引了越来越多的科学家参加。与系统生物学相关的研究论文近几年来急剧增加，2003 年的相关论文数已在 2000 年的基础上翻了一番。这与 NIH 等其他机构对此方面加大力度的资助是分不开的。目前，在英国，申请生物学相关的经费时，申请书中若无系统生物学的概念，则已很难获得基金的资助。意识到与系统生物学相关的研究论文的日益增加以及它的远大发展前景，*Nature* 已于今年 3 月专门开办了一个相关的期刊——*Molecular Systems Biology*。

近几年来，系统生物学已在美国、欧洲发展到了一定的水平，我国还在起步阶段，与美国、欧洲还有一定的差距。随着系统生物学领域的进一步发展，未来此领域对相关人才的需求将急剧增加。认识到此学科未来发展的巨大潜力，哈佛大学医学院已于 2004 年建立了世界上的第一个系统生物学系，加速培养此方面的人才。我国生物学界也开始意识到此学科的发展前景，一些单位已建立或着手准备建立系统生物学研究中心，如中国科学院上海生命科学院已与上海交通大学一起共建了系统生物学研究中心；同时，中国科技大学与中国科学院上海生命科学院已合作建立了我国第一个系统生物学人才培养基地——中国科学技术大学系统生物学系。国家科技部与中国科学院正在以上海生命科学研究院为依托单位筹建国家系统生物学实验室。相信国内这些与系统生物学相关的研究、人才培养体系的建立将为我国系统生物学的研究打下良好的基础，势必更进一步地推动我国这个研究领域的发展。

致谢：感谢南方医科大学刘靖华教授对文章提出的宝贵建议。相关的研究工作由国家自然科学基金支持。

参 考 文 献

[1] Kitano H. Systems biology: A brief overview. Science, 2002, 295: 1662—1664.

[2] Lukas J T. A signal transduction pathway model prototype I: From agonist to cellular endpoint. Biophys J, 2004, 87(3): 1406—1416.

[3] Oda K, Matsuoka Y et al. A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Systems Biol, 2005, doi:10.1038/msb4100014.

[4] Zi Z, Cho K H, Sung M H et al. In silico identification of the key components and steps in IFN-gamma induced JAK-STAT signaling pathway. FEBS Letters, 2005, 579: 1101—1108.

[5] Melkikh A V, Seleznev V D. Models of active transport of ions in biomembranes of various types of cells. J Theor Biol, 2005, 234: 403—412.

[6] Eisen M B, Spellman P T, Brown P O et al. Cluster analysis and display of genome-wide expression patterns. PNAS, 1998, 95: 14863—14868.

[7] Madeira S C, Oliveira A L. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM TCBB, 2004, 1: 24—25.

[8] Sikder D, Kodadek T. Genomic studies of transcription factor-DNA interactions. Curr Opin Chem Biol, 2005, 9: 38—45.

[9] Beer M A, Tavazoie S. Predicting gene expression from sequence. Cell, 2004, 117: 185—198.

[10] Brazma A, Schlitt T. Reverse engineering of gene regulatory networks: a finite state linear model. Genome Res, 2003, Available at http://genomebiology.com/2003/4/6/P5.

[11] Friedman N, Linial M et al. Using Bayesian networks to analyze expression data. J Comput Biol, 2000, 7(3—4): 601—620.

[12] Shmulevich I, Dougherty E R, Kim S et al. Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics, 2002, 18(10): 1319—1331.

[13] Qian J, Lin J, Luscombe N M et al. Gerstein M Prediction of regulatory networks: Genome-wide identification of transcription factor targets from gene expression data. Bioinformatics, 2003, 19: 1917—1926.

[14] Kauffman K J, Prakash P, Edwards J S. Advances in flux balance analysis. Curr Opin Biotechnol, 2003, 14: 491—496.

[15] Wang L Q, Birol I, Hatzimanikatis V. Metabolic Control Analysis under Uncertainty: Framework Development and Case Studies. Biophys J, 2004, 87: 3750—3763.

[16] Poolman M G, Venakatesh K V et al. A method for the determination of flux in elementary modes, and its application to Lactobacillus rhamnosus. Biotechnol Bioeng, 2004, 88 (5): 601—612.

[17] Alvarez-Vasquez F, Sims K J, Cowart L A et al. Simulation and validation of modelled sphingolipid metabolism in Saccharomyces cerevisiae. Nature, 2005, 433: 425—430.

[18] Shi T L, Li Y X, Cai Y D et al Computational Methods for Protein-Protein Interaction and their Application. Curr Protein & Peptide Science, 2005 (Accepted).

THE CURRENT STATE AND PERSPECTIVES OF SYSTEMS BIOLOGY

Shi Tieliu[1,2] Li Yixue[1,3]

(1 *Bioinformation Center, Shanghai Institutes for Biological Sciences, CAS, Shanghai* 200032;
2 *Bioinformation Center, Shanghai University, Shanghai* 200444;
3 *Shanghai Center for Bioinformation Technology, Shanghai* 200235)

Abstract Emerging as a new field in biology recently, systems biology provides a totally new way to study the biology processes in organisms. In order to decode the complexity of the life systematically, systems biology integrates the "-omics" and uses the high throughput methods from transcriptomics, proteomics and metabonomics to detect the dynamic activities in cell, then, it incorporates bioinformatics methods to integrate and analyze those data, and simulates the biology processes based on the model built from those integrated data. In this paper, the current state, the research field and the methods for the systems biology are introduced briefly, at the same time, some ideas about the future development for this field are also proposed.

Key words systems biology, biology system modeling, high throughput, gene regulatory network, protein-protein interaction network