

从理工科视角探索系统生物学

陈洛南¹, 王 勇², 费敏锐³, 刘曾荣¹

1. 上海大学系统生物技术研究, 上海 200444
2. 国家信息中心, 北京 100045
3. 上海大学机电工程与自动化学院, 上海 200072

[摘要] 系统生物学是在系统水平上研究生物系统, 即通过研究因生物系统组成单元之间的相互作用所组成的网络来揭示生命系统的复杂行为和功能。本文扼要介绍近期内笔者在系统生物学领域研究取得的一些进展。强调从理工科的视角出发, 首先在分子水平上结合数据与方法论的整合性观点构建生物网络, 然后用相互作用、动态行为等整体论观点来研究该网络, 达到探索在不同层面上理解和诠释复杂生命现象的目的。

[关键词] 系统生物学; 网络; 相互作用; 动态行为; 整合; 合成生物学

[中图分类号] TP391.9

[文献标识码] A

[文章编号] 1000-7857(2007)10-0005-05

Systems Biology, Its Position in Science and Technology

CHEN Luonan¹, WANG Yong², FEI Minrui³, LIU Zengrong¹

1. Institute of Systems Biology, Shanghai University, Shanghai 200444, China
2. State Information Center, Beijing 100045, China
3. College of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China

Abstract: Systems biology is the study of the interactions between the components of a biological system, and how these interactions give rise to the functions and behavior of that system. In this paper, we give a brief review on the recent studies on systems biology is given, by stressing on interactions, dynamics, networks and integration (including data sources and technologies), in particular, from the perspective of engineering viewpoint.

Key Words: systems biology; network; interaction; dynamics; integration; synthetic biology

CLC Number: TP391.9

Document Code: A

Article ID: 1000-7857(2007)10-0005-05

0 引言

后基因组时代生物学面临的一个重大挑战就是如何从整体层面上揭示生物系统中 DNA、RNA、蛋白质和各种生物小分子通过相互作用而产生生命现象。目前科学界普遍的共识是, 生命现象是不可能通过分析其组成的基本单元(基因、蛋白质和生化代谢物)而彻底解释清楚的, 需要通过从组成系统基本单元的相互作用和它们的动态行为入手来进行系统性研究, 才能最终回答物种形成和执行生命功能的科学问题。系统生物学(Systems Biology)就是在这种背景下诞生的。相对于传统生物学仅分析构成生物系统的基本单元, 系统生物学更强调在生物体的系统层面上开展对生命现象

的研究, 即以基因、蛋白质和生化反应代谢物为结点形成相互连接、具有动态特征的网络来探求和诠释各种复杂生命现象^[1-5]。

事实上, 系统生物学的整体性和整合性研究是极具生命力的。近代分子生物科学的迅速发展、基因个体分析等研究使科研工作者逐渐意识到, 物种或生命的复杂程度与基因个体数目没有直接的关系。生命从简单到复杂、从低级到高级的进化过程中, 起决定作用的不是基因个体数目, 而是生命系统中基本单元的相互作用而形成的网络复杂性。因此, 客观上需要研究者从整体性、合成的角度诠释生命系统。同时, 各种低通量、高通量的生物实验手段使得从基因到细胞、到组织、到个体的各个

收稿日期: 2007-04-15

作者简介: 陈洛南, 男, 上海市上大路 99 号上海大学系统生物技术研究, 教授, 主要从事系统生物学、系统科学等研究;

E-mail: lnchen@staff.shu.edu.cn

层次不同侧面对生命系统进行研究成为可能,利用这些手段积累的海量数据,也为研究者在整合性的观点下系统地解释生命现象提供了客观条件。

近年来,随着生物科学的进一步发展,各种高通量技术被用来系统地研究各种分子网络,例如基因芯片、酵母双杂交系统、免疫共沉淀和染色体免疫沉淀芯片(ChIP-chip)方法。通过这些实验手段得到的数据,提供了细胞或复杂生命系统的快照式轮廓,从而可以用于推断或重建蛋白质相互作用网(Protein-Protein Interaction, PPI)和基因调控网络(Gene Regulatory Network, GRN)。此外,利用来源于不同角度所积累的各种高通量数据中的信息,使获得更为完整的分子网络成为可能。进一步通过对网络整体性和动态性的更深入研究,不仅可以帮助生物学家理解复杂的生化现象,也可以从系统的广度对细胞系统的根本机制或者本质规律进行诠释。因此,这种系统生物学研究方法在短短几年中得到迅速发展,成为生物学研究中的热点。本文着重从理工科角度出发,介绍笔者最近在系统生物学领域的一些研究工作。通过一些例子,强调以网络、相互作用、动态行为等整体论观点,并结合数据与方法论的整合性观点,探索如何对复杂生命现象进行不同层面的理解和诠释。

1 单细胞内各类分子生物网和系统研究

动力学行为以不同的层次广泛存在于生命体中。无论是从理论角度还是生物实验角度,复杂生物动态系统的建模、分析和预测都对现代科学形成了巨大挑战。截止到目前,得到最深入广泛研究的动态行为之一就是24 h周期节律现象,普遍认为这是一种在分子水平上由基因调控负反馈环所形成的极限环产生的周期振动。

就单细胞而言,其内存在着不同形式的网络,比如基因调控网、蛋白质作用网和代谢网。笔者工作主要集中在基因调控网的研究上,尤其是关于基因调控网的设计。随着描述基因调控机制的数学工具和实验手段的长足进步,急需建立起从系统层面阐明单细胞内动力学行为的更成熟的理论模型和更通用的技术手段。于是合成生物学作为一个与系统生物学紧密相关的新研究领域诞生了。它结合科学概念和工程实践,旨在利用系统生物学的概念、技术来设计和建立具有全新生命功能的元件和系统。合成生物学试图由合成的角度看待生物学,结合分子生物学、工程和数学各领域的方法来设计和实现全新的细胞和生命行为。基因工程的最新进展已经使得设计和实现人工或者合成的基因调控元件和网络成为现实,这一点从理论和生物实验角度都得到了很好的证实。事实上,一些较为简单的基因调控网络,例如基因转换开关和基因振动物子,都是首先从理论上设计并证明

其功能是能够实现的,然后由生物学家设计生物实验在生物体内得到成功验证。这些简单基因调控元件的设计成功,标志着在DNA层次上实现对生命过程的调节和逻辑操控迈出了可喜的第一步,也是极为重要的一步,因为它们不仅可以作为构建和合成更为复杂人工生命的基本单元,也对未来在实际应用中实现生物技术和基因治疗起着非常关键的作用^[6-7]。从工程的角度看,这样的研究是一种正向工程研究方式^[8-13]。在这个领域,笔者提出了一系列更为通用的方法用于构建单细胞内的基因调控网,并发展了相应的理论技术来设计生物分子网络。

与单细胞内的网络和系统有关的主要工作概括如下。

1) 基于单调动力系统的基因开关和基因振动物子设计方法和理论^[8-10, 14]。

2) 通过探索网络结构中的正反馈和负反馈环来设计通用的人工合成基因调控网络^[11-12]。

3) 基因调控网络建模^[15-17],以及基于非线性动力理论和控制理论(LMI or Lure systems)分析其动态行为(稳定性和分岔现象)。

4) 考虑生物体内细胞周期影响的生物分子网络建模的统一通用框架和理论^[18]。

从非线性动力学的观点来看,分析大规模生命系统主要有3个较为本质的困难,即非线性、噪声和时滞。在以上工作中,笔者主要是利用细胞系统的特殊性质来克服这些非线性和时滞现象的困难。下面专门介绍针对如何处理生命系统中的噪声问题所做的一些工作。

2 多细胞系统的动态协作机理研究

协作行为对于细胞群体通过细胞通信来实现对外部信息交换的响应是非常重要的,这种机制广泛存在于原核和真核生物中。细胞间的通信或者合作能力对于在一定环境下确保生命体各层次细胞活动以一种精准和可靠的方式运作至关重要,所以深入理解分子间的协作行为(如趋药性和群体感应)发生的确切机制,也是系统生物学的重要研究方向。而探索这一方向的先决条件就是数学和生物学的结合。

一般来讲,细胞群的协作行为是通过细胞内的信息在细胞间的反应传递来完成的。细胞间的反应传递是通过细胞内信号传导到相邻细胞,然后这些局部的信号在分子、组织、器官和身体层次进一步集成而产生全局的细胞响应。

这方面研究的典型代表是噪声在细胞群协作行为的研究。事实上,所有的细胞组分都会具有两种噪声:由细胞内的个体分子的随机生成或者降解而产生细胞内

的噪声,由外部环境的摄动变化带来的细胞外噪声。所以,此时基因间的调控关系本质上就是一种随机噪声过程。研究这个过程既要考虑到由于每个细胞内较低的分子数量造成的随机摄动,又要考虑到来源于外部环境的不确定性。这样的随机噪声既会影响整个系统的动力学行为,同时又会被生物体正面合理地利用来产生一些积极的生物功能,如实现细胞间的同步和通信等。对于这个复杂的现象,笔者已经做到了从理论上构建模型,同时发展了一套数值算法来分析一群单细胞之间由于噪声引起的协作行为。在这些研究工作中,不但考虑了随机摄动的影响,也考虑了信号在生物体内扩散的过程。所有的工作都是基于确定性或者随机的动力学分析,揭示了噪声在生物体中起着非常重要的作用,特别是它的积极调节协作动力学的行为^[19]。

与多细胞系统有关的主要研究结果总结如下。

- 1) 在有耦合和无耦合情况下,通过信号分子实现多细胞系统的基因振子同步^[13, 20]。
- 2) 研究了有噪声和无噪声情况下耦合非一致细胞的协作行为^[21-22],以及生物同步中的瞬态复位机制^[23]。
- 3) 通过外部输入,研究多细胞系统的同步控制策略^[24]。
- 4) 基于主方程、累积方程和随机模拟,分析由细胞外噪声诱导的随机同步分子通信机制^[25-26]。

3 整合与系统

在系统生物学研究中,除了上述用网络通过动力学研究来解释生命现象外,另一个重要任务是如何整合大量的生物数据来确定所需要研究的网络。基因调控关系预测的研究是重要方面。

大量积累的微阵列(microarray)基因表达数据为从生物实验数据中学习基因间调控关系,进而构建基因调控网络和信号传导通路提供了新的途径。其数据的特点是一次微阵列实验能获得细胞在某一条件下的全基因组表达数据,包含成千上万的基因在细胞中的相对或绝对丰度。但是由于生物实验条件所限,一次实验只能获得少数的样本点,故基因数据维数远远高于样本个数,这就导致了由于基因表达数据的“维数灾难”所造成的数据相对数学模型不足的困难,使得集成各种不同的信息源、从实验数据出发通过反向工程方法重构可靠的分子网络成为最富挑战性的问题。

一种新的重建基因调控网络方法——GRNInfer^[27],与传统的只依靠单一时间序列数据集来推断基因调控网的方法不同。它的特点是能够在统一的数学规划框架下系统地集成多个时间序列数据集,并且允许这些数据集是在不同条件(细胞周期的不同阶段、药物作用时间、肿瘤类型、不同病人等)下的全基因组表达数据。

这种方法在理论上保证从所有数据集中可以得到最相容的网络结构,因此极大地缓解了基因调控网络重建方法对微阵列数据样本点个数少所造成的困难,并且显著地提高了基因调控关系预测的可靠度和精度。GRNInfer也可以进一步扩展有机集成或整合其他已知信息,如文献[27]中发表的基因调控关系、基因共表达关系、ChIP-chip 数据;识别对外部环境(温度、化合物或者药物)有显著响应的基因或者是信号传导通路中的信号受体;检测基于多数据构建网络中的单物种内或多物种间的富有进化意义的保守模式或模块。

4 生物分子相互作用及分子网络比对

高通量生物实验手段得到的数据通常有两个主要问题,一是相对的数量和信息量不足,只能观测到复杂生命体中的一个部分;二是数据的质量不够可靠,通常有大量的假阳性数据存在。目前已经陆续有不少计算方法针对现有基因组学、蛋白质组学数据的这些特点被用来推断或重建生物分子网络。

4.1 蛋白相互作用研究

近年来,一类非常有趣的基于统计分析的算法被用于从结构域层次发现潜在的新的蛋白质相互作用。这类方法首先假设蛋白质的结构域是构成蛋白质并且形成相互作用的基本单元,并进一步假设每一对蛋白质相互作用至少是由于一对结构域相互作用导致的;然后以高通量生物实验观测到的蛋白相互作用作为真实的训练数据,通过首先计算每一对结构域相互作用的可能概率,再利用计算得到的所有结构域相互作用信息,来预测新的蛋白质相互作用。针对蛋白相互作用这个问题,从生物角度讲,结构域作为生命进化过程中蛋白质的保守单元,具有很清晰的生物意义,所以可以得到较为可信的蛋白相互作用。基于此原理,笔者研究了高通量蛋白质相互作用数据和蛋白质与其结构域组成的关系,开发了一种新的关联概率方法 APM (Association Probabilistic Method)^[28]。该方法是利用蛋白质的结构域组成来预测蛋白质相互作用。APM 在预测精度和计算效率方面均超过了目前已有的方法,而且 APM 也可以通过一个简单变化扩展应用到多物种的蛋白相互作用数据上^[7, 28]。

4.2 生物分子网络比对研究

分子网络在细胞中以一种相当精准的控制方式执行着复杂的生命功能。不同物种间的差异,不仅是因为组成蛋白个数上的差异,也是因为其分子网络在相互作用和层次上的区别。因此,极有必要通过比较它们的分子网络的区别和联系来解释物种进化的奥秘。而且好的网络比对算法可以直接应用于比较分析信号传导通路、不同物种分子网络间的保守区域,并能发现新的生物功

能或者为理解蛋白质相互作用的进化提供新的证据。有的学者甚至对比了网络比对方法和序列比对方法的发展历程,认为在系统生物学时代需要发展类似于序列比对 BLAST 那样的理论,并进一步应用于网络比对的计算。

从本质来讲,网络比对问题就是要通过比较分子生物网络来探测同物种内或者跨物种中的保守子结构或子网络。因为网络比对问题既要考虑结点相似性,同时也要考虑边的相似性,具有很高的计算复杂性,所以目前通用的方法都是针对一类具体的有特殊结构的网络,例如仅考虑无向图或不含圈的通路,采用启发式的算法来缓解对大规模计算量的要求。针对这个由于计算复杂性带来的难点,笔者开发了一个新的网络比对方法——MNAAligner^[29]。该方法用数学规划中的整数二次规划模型更精确地刻画两个网络的相似性。且 MNAAligner 是一个相当通用的方法,既可以被广泛的用在有向、无向网络比对中,也可以用在无权、赋权网络中^[7, 30]。

5 计算系统生物学的研究

计算系统生物学的研究重点主要体现在基因组学(Genomics)和蛋白组学(Proteomics)两方面,用数学、信息学和计算机科学的方法和技术来系统地解决生物问题。计算系统生物学中的很多研究内容常常与计算生物学或生物信息学有交叉。

在这一领域,笔者构建了一系列数学模型并设计了算法来解决实际的生物问题,可以简单地总结如下。

1) 基于数学规划模型从 SNP 数据研究单体型组装问题^[30-31]。

2) 基于图论理论研究蛋白质序列在三角网格上的唯一最优折叠构象^[32-33]。

3) 基于多目标规划理论和算法的蛋白质比较研究,用以从蛋白质三维结构数据解释趋异进化、识别循环排列(circular permutation)区域以及检测活性位点区域^[34-35]。

4) 基于数学规划模型的蛋白质多结构比对算法研究^[36]。

5) 基于统计分析的从质谱数据识别蛋白质^[37]。

6) 基于系统生物学观点的基因网络模式结构的信号处理机制研究^[38]。

7) 基于几何拓扑特征的蛋白质结构比较方法研究^[39],及其在外膜蛋白分类中的应用^[40]。

6 结论

系统生物学的根本挑战在于建立完整的、细致的生物分子间联系描述,并藉此研究分子相互作用与生理响应之间的紧密关联。本文通过系统性研究组成复杂生

物系统,如代谢通路、细胞器官、细胞、生理系统和生物体的各种单元之间的关系和连接所建立的网络,试图从理工科的视角和高度(如正向工程和负向工程的方式),最终对生命系统建立一个可理解的模型,它将是理解生命现象和应用到人类疾病治疗的一把钥匙。目前这方面工作刚刚开始,有大量工作等待人们去开发和研究。

另外,本文特别想通过强调系统生物学具有对理论的依赖和建立模型的需求的特点,即从理工科角度看待系统生物学,来构建一个尽可能接近真正生物系统的理论模型。这也是系统生物学研究的终极目标。可以说理工科的思想和方法论将会始终贯穿在系统生物学研究的每一个阶段。从这个意义上讲,系统生物学开创了生物学和其他学科进行大交叉研究的一个全新时代。

致谢:本文的工作主要是与中国科学院数学与系统科学研究院的章祥荪研究员的研究小组、东京大学生产技术研究所的合原一幸教授研究小组合作完成的,特此致谢。

参考文献(References)

- [1] HOOD L. A personal view of molecular technology and how it has changed biology [J]. J Proteome Res, 2002, 1 (5): 399-409.
- [2] IDEKER T, *et al.* Integrated genomics and proteomic analyses of a systematically perturbed metabolic network [J]. Science, 2001, 292: 929-934.
- [3] NOBLE D. Modeling the heart—from genes to cells to the whole organ[J]. Science, 2002, 295: 1678-1682.
- [4] ARBEITMAN M N, *et al.* Gene expression during the life cycle of *Drosophila melanogaster* [J]. Science, 2002, 297: 2270.
- [5] KITANO H. Systems biology: A brief overview[J]. Science, 2002, 295: 1662-1664.
- [6] CHARUSANTI P, HU X, CHEN L, *et al.* A mathematical model of BCR-ABL autophosphorylation, signaling through the CRKL pathway, and gleevec dynamics in chronic myeloid leukemia [J]. Discrete and Continuous Dynamical Systems—Series B, 2004, 4: 99-114.
- [7] ZHANG S, JIN G, ZHANG X S, *et al.* Discovering functions and revealing mechanisms at molecular level from biological networks[J]. Proteomics, in press, 2007.
- [8] WANG R, JING Z, CHEN L. Modelling periodic oscillation in gene regulatory networks by cyclic feedback systems[J]. Bulletin of Mathematical Biology, 2005, 67: 339-367.
- [9] WANG R, CHEN L, AIHARA K. Construction of genetic oscillators with interlocked feedback networks [J]. Journal of Theoretical Biology, 2006, 242: 454-463.
- [10] WANG R, CHEN L, AIHARA K. Detection of cellular rhythms and global stability within interlocked feedback systems[J]. Mathematical Biosciences, in press, 2007.

- [11] WANG R, ZHOU T, JING Z, *et al.* Modelling periodic oscillation of biological systems with multiple time scale networks[J]. *Systems Biology*, 2004, 1: 71–84.
- [12] CHEN L, WANG R. Designing gene regulatory networks with specified functions[J]. *IEEE Trans on Circuits and Systems-I*, 2006, 53: 2444–2450.
- [13] WANG R, CHEN L. Synchronizing genetic oscillators by signalling molecules [J]. *Journal of Biological Rhythms*, 2005, 20: 257–269.
- [14] KOBAYASHI T, CHEN L, AIHARA K. Modeling genetic switches with positive feedback loops [J]. *Journal of Theoretical Biology*, 2002, 221: 379–399.
- [15] CHEN L, AIHARA, K. Stability of genetic regulatory networks with time delay [J]. *IEEE Trans. on Circuits and Systems-I*, 2002, 49: 602–608.
- [16] CHEN L, AIHARA K. A Model of periodic oscillation for genetic regulatory systems [J]. *IEEE Trans on Circuits and Systems-I*, 2002, 49: 1429–1436.
- [17] LI C, CHEN L, AIHARA K. Stability of genetic networks with SUM regulatory logic[J]. *IEEE Trans on Circuits and Systems-I*, 2006, 53: 2451–2458.
- [18] CHEN L, WANG R, KOBAYASHI T, *et al.* Dynamics of gene regulatory networks with cell division cycle [J]. *Physical Review E*, 2004, 70: 011909.
- [19] SPRINGER M, PAULSSON J. Harmonies from noise[J]. *Nature*, 2006, 439: 27–28.
- [20] ZHOU T, CHEN L, WANG R. A mechanism of synchronization in interacting multicell genetic systems [J]. *Physica D*, 2005, 211: 107–127.
- [21] LI C, CHEN L, AIHARA K. Synchronization of coupled nonidentical genetic oscillators [J]. *Physical Biology*, 2006, 3: 37–44.
- [22] LI C, CHEN L, AIHARA K. Stochastic synchronization of genetic oscillator networks[J/OL]. *BMC Systems Biology*, [2007]. <http://www.biomedcentral.com/1752-0509/1/6>.
- [23] LI C, CHEN L, AIHARA K. Transient Resetting: a novel mechanism for biological synchrony [J/OL]. *PLoS Computational Biology*, [2006]. http://compbiol.plosjournals.org/perlserv/?request=get_document&doi=10.1371%2Fjournal.pcbi.0020103.
- [24] WANG R, CHEN L, AIHARA K. Synchronizing a multicellular system by external input: An artificial control strategy[J]. *Bioinformatics*, 2006, 22: 1775–1781.
- [25] CHEN L, WANG R, ZHOU T, *et al.* Noise-induced cooperative behavior in a multi-cell system [J]. *Bioinformatics*, 2005, 21: 2722–2729.
- [26] ZHOU T, CHEN L, AIHARA K. Molecular communication through Stochastic Synchronization Induced by Extracellular Fluctuations[J]. *Phys Rev Lett*, 2005, 95: 178103.
- [27] WANG Y, JOSHI T, ZHANG X S, *et al.* Inferring gene regulatory networks from multiple microarray datasets[J]. *Bioinformatics*, 2006, 22: 2413–2420.
- [28] CHEN L, WU L Y, WANG Y, *et al.* Inferring protein interactions from experimental data by association probabilistic method[J]. *Proteins*, 2006, 62: 833–837.
- [29] LI Z, ZHANG S, WANG Y, *et al.* Alignment of molecular networks by integer quadratic programming [J]. *Bioinformatics*, in press, 2007.
- [30] LI Z, ZHOU W, ZHANG X S, *et al.* A parsimonious tree-grow method for haplotype inference [J]. *bioinformatics*, 2005, 21: 3475–3481.
- [31] ZHANG X S, WANG R S, WU L Y, *et al.* Models and algorithms for haplotyping problem [J]. *Current Bioinformatics*, 2006, 1: 105–114.
- [32] LI Z, ZHANG X S, CHEN L. Unique optimal foldings of protein on a triangular lattice [J]. *Journal of Applied Bioinformatics*, 2005, 4: 105–116.
- [33] ZHANG X, WANG Y, ZHAN Z, *et al.* Exploring protein optimal HP configurations by self-organizing mapping [J]. *J Bioinformatics and Comput Biol*, 2005, 3: 385–400.
- [34] CHEN L, ZHOU T, TANG Y. Protein structure alignment by deterministic annealing [J]. *Bioinformatics*, 2005, 21: 51–62.
- [35] CHEN L, WU L Y, WANG Y, *et al.* Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison[J/OL]. *BMC Structural Biology*, [2006]. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1574323>.
- [36] ZHOU T, CHEN L, TANG Y, *et al.* Aligning multiple protein structures by deterministic annealing [J]. *Journal of Bioinformatics and Computational Biology*, 2005, 3: 837–860.
- [37] SONG Z, CHEN L, GANAPATHY A, *et al.* Development and assessment of scoring functions for protein identification using peptide mass fingerprinting data [J]. *Electrophoresis*, 2007, 28: 864–870.
- [38] LI C, CHEN L, AIHARA K. A systems biology perspective on signal processing in genetic network motifs[J]. *IEEE Signal Processing Magazine*, 2007, 24: 136–147.
- [39] WU Z, WANG Y, FENG E, *et al.* A new geometric-topological method to measure protein fold similarity [J]. *Chemical Physics Letters*, 2007, 433: 432–438.
- [40] WU Z, FENG E, WANG Y, *et al.* Discrimination of outer membrane proteins by a new measure of information discrepancy [J]. *Protein and Peptide Letters*, 2007, 14: 37–44.

(责任编辑 赵佳)