以信息系统的观点了解基因组

李衍达
(清华大学生物信息研究所, 智能技术与系统国家重点实验室, 北京 100084)

摘 要: 后基因组时代的核心任务是了解基因组与蛋白质组的功能. 由于生物是在遗传信息与外界信息作用下的一个复杂、有序的动态系统, 基因组包含生命所需的信息, 包括其产物和调控的信息, 所以, 本质上, 基因组是一个信息系统. 基因组的调控与信息系统的调控具有相同的规律. 据此, 本文提出将信息与信息结构作为内涵对基因组序列作为整体系统分析的设想, 以信息系统的观点了解基因组.

关键词: 基因组; 蛋白质组; 复杂性; 信息结构; 信息系统
中图分类号: Q522 文献标识码: A 文章编号: 0372-2112 (2001) 12A-1731-04

Understanding Genome from Information System Point of View

LI Yan-da
(*Bioinformatics Institute, Tsinghua University, State Key Laboratory of Intelligent Technology and System, Beijing* 100084, *China*)

Abstract: To understand the function genome and proteome is the main task of post genome era. Biosome is somewhat a complex, ordinal system which is excited both by genetic and environmental information. Genome has contained all the genetic information needed, including its products and regulation information. In nature, genome is an information system, which obeys the same regulation law as a general information system does. Consequently, we have to understand the function genome from the information system point of view, and treat biological signal as the core content of this scheme.

Key words: genome; proteome; complexity; information structure; information system

1 后基因组时代对我们提出的要求

人类基因组计划的初步完成, 宣告了后基因组时代的到来. 随着HGP计划的进行, 核酸与蛋白质的数据成指数增长(如图1所示). 大量的生物数据等待人们去分析, 生物信息学成为研究的焦点. 目前, 我们对数据的分析、解释远远落后于实验的速度. 因而急需应用数学、信息科学、化学、计算机科学来分析和理解数据, 以便更深入地理解生命现象.

后基因组时代的核心任务是了解基因组与蛋白质组的功能[1], 或者说是解读遗传密码. 这是揭示生命奥秘的关键的一步. 通过大量的实践, 人们认识到以往一个基因、一个蛋白质的研究并不能解释基因的功能. 必须在众多基因与蛋白质的相互作用中了解其功能. 这使得分子生物学研究的思路产生了转变(如图2所示), 即更注重从系统的角度, 认识基因之间, 蛋白质之间, 不同转录、表达现象之间的联系, 以及从这种联系构成的系统整体特征去了解其功能.

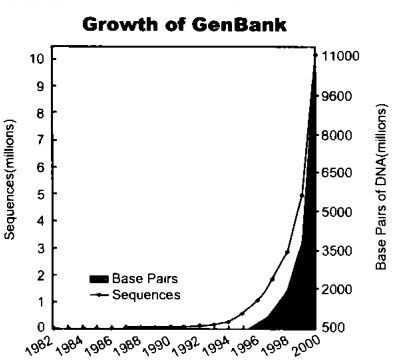

图 1

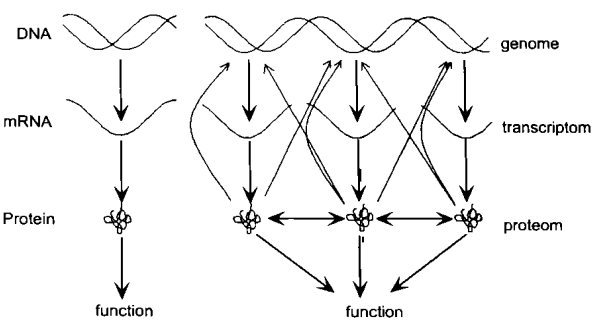

图 2 分子生物学研究思路的转变

我们知道细胞是揭示生命奥秘的基础, 细胞研究的三个基本问题是:

(1) 细胞的基因组如何在时间、空间上有序的表达;

(2) 细胞基因组表达的产物——结构蛋白质、核酸、脂与多糖等是如何逐级装配成各种细胞器与组织的, 自组装的调控程序与机理是什么;

(3) 活性分子, 信号分子是如何调节生命过程的. 包括染

收稿日期: 2001-06-27; 修回日期: 2001-09-15
基金项目: 国家自然科学基金(No. 69872018, 69935020), 北京市科技合同项目(No. 9550213000), 清华大学基础和创新研究基金.

色体、DNA 与蛋白质如何相互作用, 细胞信号传递过程与路径的分析等.

基因组, 蛋白质组功能的研究与上述三个基本问题是密切相关的, 而进行上述问题的研究需要对 DNA 整体序列的理解.

2 生物信息学研究的特点

生物信息学研究的根本目标是揭示基因组信息结构的复杂性及遗传语言的根本规律. 因而基因组的信息结构及因其结构的复杂性[2] 而带来对基因组性能的影响成为生物信息学研究的重要内容. 可以认为后基因组时代对生物信息学研究的要求产生了新的变化, 这就是:

- 从单因素的影响到多因素影响的研究;
- 从数据的分类、统计到对编码、语法的解读;
- 从偶然、个别的实验结果得到新的知识到对生物内部活动规律的理解;
- 从单个生物分子到多种生物分子相互作用的分析.

以上内容, 无不要求对生物有内在规律性的理解, 而且都离不开将生物信号与信息作为内涵来理解.

由于生物是在遗传信息与外界信息作用下的一个复杂、有序的动态系统. 基因组包含生命所需的信息, 包括其产物和调控的信息. 所以, 本质上, 基因组是一个信息系统. 基因组的调控与信息系统的调控具有相同的规律. 因此, 现有的一些研究方法显出以下的一些不足:

(1) 把基因组序列作为一个整体系统来分析不够;
(2) 把物质结构与信息结构联系分析不够;
(3) 把活性与信息结构联系分析不够;
(4) 把信息结构作为动力系统的分析不够;
(5) 作为一种遗传语言的分析不够.

据此, 我们提出将信息与信息结构作为内涵对基因组序列作为整体系统分析的设想, 以信息系统的观点了解基因组. 应该指出, 根据复杂系统理论, 我们不仅要看到组成系统的基本单元的数量, 更要看到单元之间的相互作用, 会产生新的性能. 可能复杂的信息系统正是我们理解基因组及其功能的关键.

3 基因组信息结构的复杂性

通过对全基因组 DNA 序列的分析, 可以说明基因组信息结构的复杂性. 我们对 E. Coli 全基因组腺嘌呤 100bp 浓度分布进行的分析, 说明具有复杂系统的自相似性– – 具长程相关性(如图 3[3] 所示) .

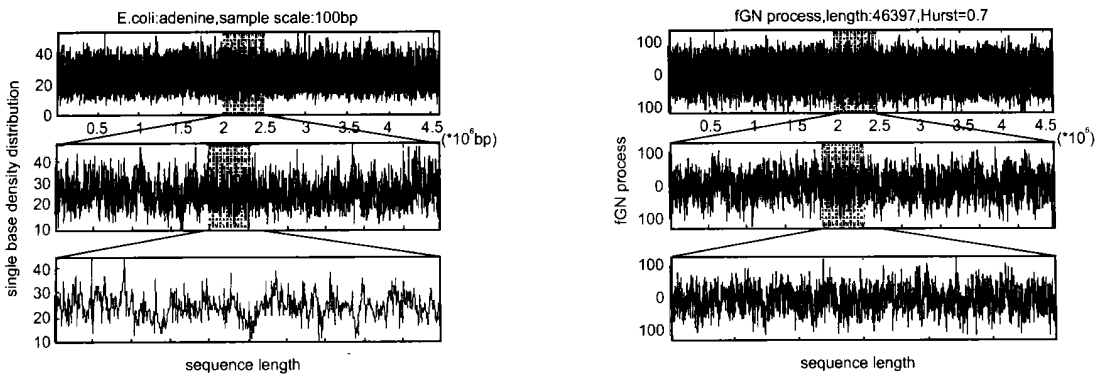

图 3 E. coli 腺嘌呤 100bp 浓度分布曲线及一个同等长度 $H = 0.7$ 的 fGN 过程.

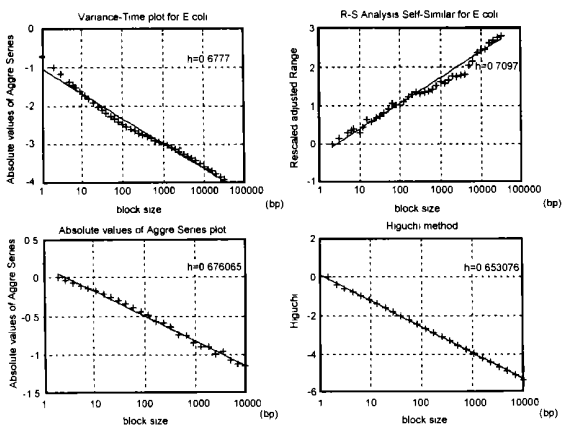

图 4 用各种方法估计 E. coli 腺嘌呤浓度分布 Hurst 系数的拟合直线图

上图的左方是 E. Coli 全基因组腺嘌呤 100bp 浓度分布图, 右方是一个典型的分形高斯过程. 从直观可以看出, 它们很相似, 可能具有长程相关性. 为了证明全基因组 DNA 序列具有复杂系统的自相似性, 我们对左图用各种方法估计其 Hurst 指数, 结果如图 4 所示. 结果表明, 其 Hurst 指数均大于 0.5, 存在自相似特性.

通过进一步的研究, 发现利用 DNA 序列的分维数可以对外显子、内含子进行分类. 因此, 自相似性的确是 DNA 序列的一个重要特征. 但是, DNA 序列信息结构的自相似性的生物含义及其起源, 尚有争议. 有人(Grosberg 等) 认为长程相关性与 DNA 分子的高级结构有关; 有人认为(Voss, 1992) 自相似性为 DNA 的复制和功能提供一种尺度无关的纠错能力; Li(Li, 1989, 1991) 等认为自相似性是由于 DNA 进化采用扩增– 突变模型的结果.

另外, 应该指出生物活性与复杂系统性质有密切关系. 如果没有复杂系统的突现概念就难以理解活性. 因为按线性关系, 从无活性不能得出活性. 只有非线性关系, 即复杂系统有其组成部分没有的性质: 部分之和小于整体的概念, 才能出现

活性.

4 基因组的调控与信息系统的调控

了解基因组调控机理是了解基因组功能的基础, 那么, 基因组调控的机理与信息系统的调控原理是否一致? 下面, 我们举例说明生物分子的局部调控与信息系统的调控是一致的[4, 5], 两者可以对比. 例如:

1 糖原代谢控制与信息系统的分支控制

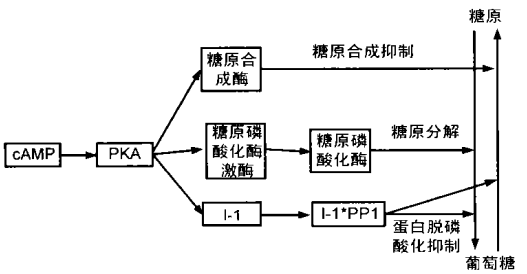

图 5 以糖原代谢的控制为代表的分支控制

2 组蛋白的表达与负反馈控制

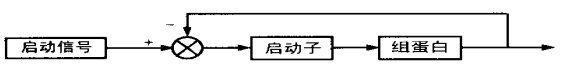

图 6 以组蛋白表达控制为代表的负反馈控制

3 胰岛素受体自身磷酸化与正反馈控制

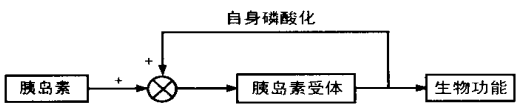

图 7 以胰岛素受体自身磷酸化作用为代表的简单正反馈控制

4 细胞周期素控制与复杂正、负反馈控制

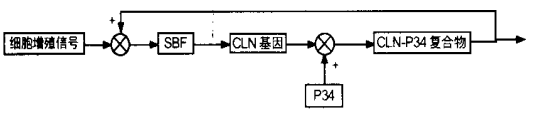

图 8 以细胞周期素控制为代表的复杂正反馈控制

5 Ras 蛋白活性控制与延时控制

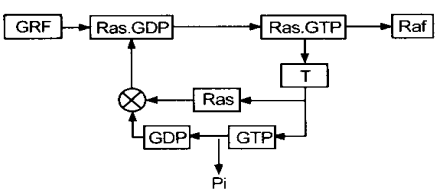

图 9 以 Ras 蛋白活性控制为代表的延时控制

由以上可得出结论:

(1) 基因组调控符合一般的控制模型.

(2) 控制模型的控制特性由代表其相互作用及其本身特性的信号系统决定.

从另一角度的来看, 我们利用符合生物分子调控机理的信号系统来模拟生物的调控过程, 也可以产生类似的生物特性. 例如, 钙波是细胞内一个重要的信号, 我们可以用钙波的信号系统模型描述钙波特性[6].

人体内钙波(Ca^{2+})有多种的振荡特性, 钉形波形是其中一种, 其特点是振荡频率随外激励改变, 而振荡幅度基本不变. 我们根据细胞内对钙波调控的两种机制, 即通过内质网、质膜钙泵调控的快系统与通过线粒体、质膜 Na^{+} —Ca^{2+} 交换器的慢系统给出单细胞钙波的控制模型如下:

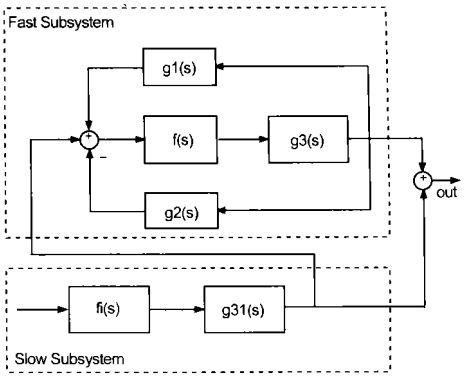

图 10 单细胞钙波的信号系统模型

适当调节上模型的参数, 即可得出类似钉形的振荡波形, 在不同激励强度下, 频率随激励强度上升, 而振荡幅度基本不变(如图 11 所示).

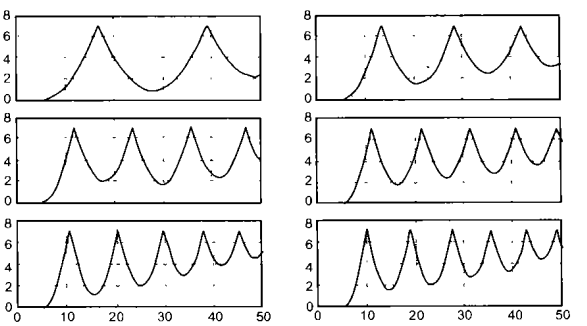

图 11 模型仿真得到的钙波震荡波形

因此, 可以说基因组的信息系统(或信号系统)决定基因组的调控性能.

5 几点结论

(1)基因组、蛋白质组的研究从测序转向功能研究, 从实验数据转向理解数据, 寻找基本规律.

(2)在重视实验、结构分析的基础上, 生物信息学应重视生物内信息系统的分析.

信息系统的观点主要是把信号、活性、进化与调控四个方面在信息系统内统一起来, 形成一个整体系统.

(3)以信息系统观点了解基因组主要有以下几个方面的特点:

(a)以整个系统的信号(信息为其内涵)的调控为主要线索;

(b)DNA, RNA, 蛋白质都有其信号的表示方式;

(c)活性分子是一类特殊的信号系统, 它以自组装, 自调控作为其特征;

(d)生物的进化既受活性的约束, 又必须符合信息系统作为一类非线性动力系统的演化规律;

(e)生物内部的调控作用是信号(信息)的调控过程;

(f)把环境刺激及其响应看作一种信号的刺激和信号系统的响应来看待.

由此,把信号、活性、进化、调控的内在规律统一在信息系统中.

(4)生物信息的研究可从局部与整体两方面入手.

局部:基因、蛋白质及其内部的信息结构,如基因识别、功能位点预测、基因功能预测、蛋白质结构分类与预测等

整体:基因组信息结构,基因组比较,基因表达及其调控等.

但是,无论从局部还是整体,都应注重其信息的表示,信号的识别与相互作用.都应从系统的角度,特别是复杂系统的角度分析.信号系统、信息结构、复杂性这三者是我们应关注的重点.

(5)生物信息学工作中有一部分需依赖于强大的计算能力,如比对、测序、SNP、生物芯片数据分析、可视化与仿真,等等,这部分十分重要,并将发挥极大威力.但是,有一部分的研究,仅有强大的计算能力是不够的,需依赖分子生物学与信息、数学、系统论的结合,创造性的找出有效的分析与理解方法.其中,信息系统的研究是关键性的.后者的成功,将开辟出新的道路:基因组的功能研究,蛋白质组研究,基因组的调控、信号传导过程研究,多基因病的分析,以及疾病诊断,药物设计.其中,相当大一部分问题的解决取决于基因组信息系统的研究工作的突破.

参考文献:

[1] Lu Xin, et al. Characterizing self-similarity in bacteria DNA sequence [J]. Physical Review E, 1998, 58(3): 3578.

[2] Lu Xin, et al. Typical control element in gene regulation pathway [A]. Biotechmology 2000 [C], Berlin, Sep., 2000.

[3] Cai Jun, et al. Modeling of molecular regulation networks with metar model [A]. Second International Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk [C], Russia, Aug. 7, 2000.

[4] Field S. The future is function [J]. Nature Genetics, 1997, 15: 325-327.

[5] Weng G, et al. Complexity in biological signaling systems [J]. Science, 1999, 284: 92-96.

[6] Wiener N. Cybernetics, or control and communication in the animal and the machine [M]. The MIT Press and John Wiley Sons, Inc. 1961.

[7] Ideker T, et al. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network [J]. Science, 2001 May 4; 292: 929-934.

作者简介:

李衍达 中国科学院院士,清华大学自动化系教授,博士生导师.1936年10月出生于广东省东莞市.现任国务院学位委员会委员、北京生物工程学会生物信息学专业委员会主任、IEEE高级会员、中国自然科学基金委员会委员.主要学术方向为:信号处理,生物信息学与智能信息处理.已发表论文百余篇及多部著作.