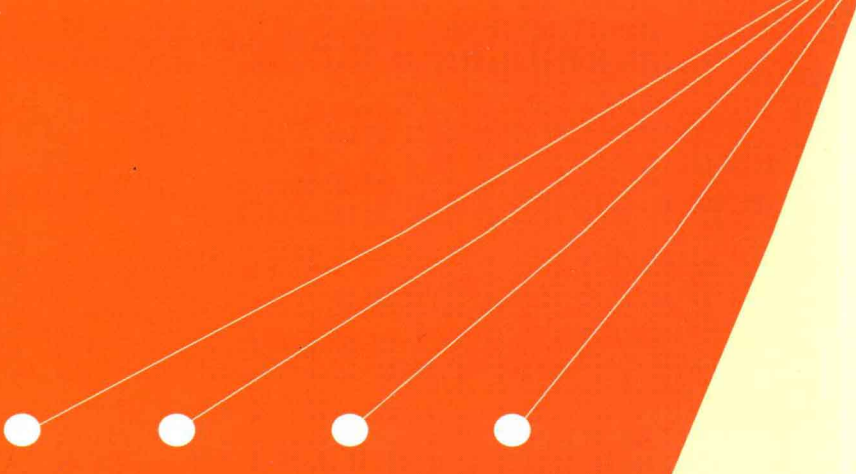
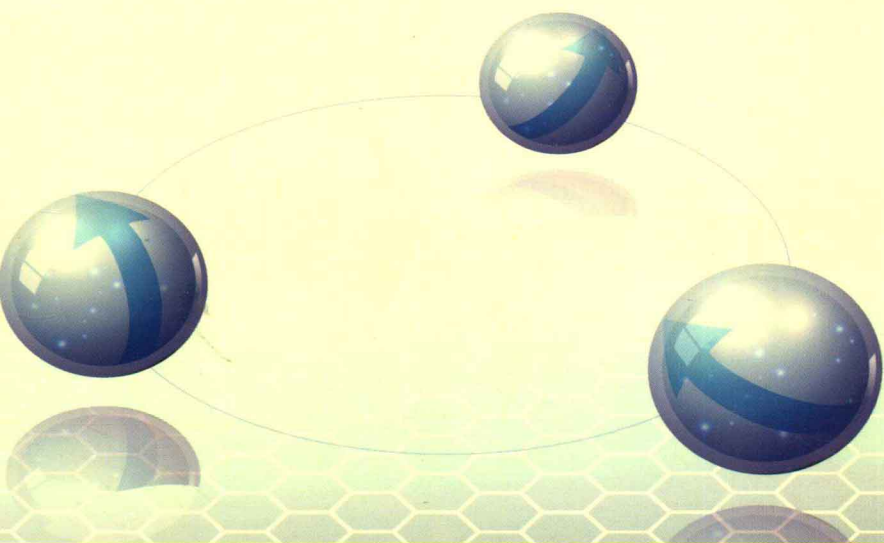




局部基因

孟良玉 著

网络技术及其应用



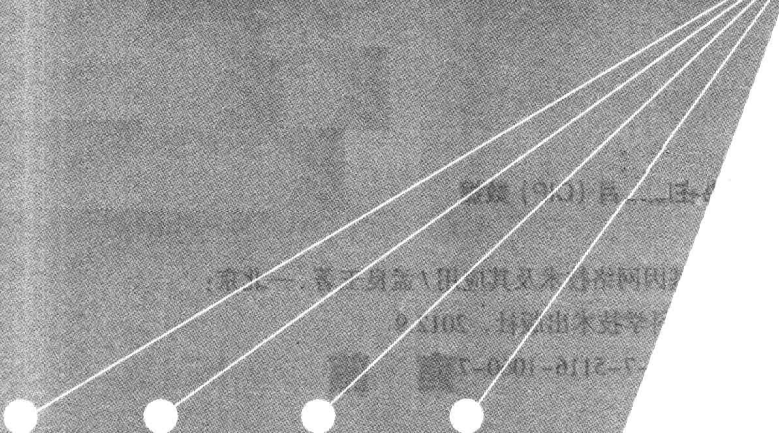
中国农业科学技术出版社

局域网

第2版

网络技术及其应用

中国铁道出版社



局部基因

孟良玉 著

网络技术及其应用

中国农业科学技术出版社

图书在版编目 (CIP) 数据

局部基因网络技术及其应用 / 孟良玉著. —北京:
中国农业科学技术出版社, 2012.9
ISBN 978-7-5116-1060-7

I. ①局… II. ①孟… III. ①疾病—基因—关系—营
养代谢—研究 IV. ①R394 ②Q591.1

中国版本图书馆 CIP 数据核字 (2012) 第 202417 号

责任编辑 张孝安 白姗姗
责任校对 贾晓红 郭苗苗

出 版 者 中国农业科学技术出版社
北京市中关村南大街 12 号 邮编: 100081
电 话 (010) 82109708 (编辑室) (010) 82109704 (发行部)
(010) 82109703 (读者服务部)
传 真 (010) 82109708
网 址 <http://www.castp.cn>
经 销 者 新华书店北京发行所
印 刷 者 北京富泰印刷有限责任公司
开 本 880mm × 1.230mm 1/32
印 张 3.875
字 数 84 千字
版 次 2012 年 9 月第 1 版 2012 年 9 月第 1 次印刷
定 价 20.00 元

❧ 版权所有 · 侵权必究 ❧

前 言

复杂疾病一般对应着一群直接相关的基因，目前，还非常缺乏有效地认识多个基因之间的互相作用及基因与营养代谢之间关系的方法，短期内也很难弄清楚它们在复杂的细胞网络中的作用方式。随着人类全基因组测序工作的完成以及相应的功能基因组学乃至系统生物学研究手段的建立，产生了大量功能基因数据，在此基础上建立一批较为可行的计算或分析方法用以处理多基因之间及其与营养代谢间的复杂关系，已经具有较高的可行性。

本书介绍的内容就是在相关技术和学科发展的基础上提出的，应用局部基因网络技术研究了乳腺癌、高血压、肥胖等复杂疾病的易感基因（网络）与营养代谢之间的关系，从分子（基因）水平上解释疾病与代谢间关系的复杂性，并从营养代谢的角度提出或解释疾病的有效的预防或治疗措施。它为研究复杂疾病提供一个新思路，尝试应用一种新的方法来研究疾病的致病机理，为探讨这类疾病的病因、制订防治措施、提供营养学建议作了一定的尝试，进而试图为开展疾病的预防和日常饮食提供理论依据，有针对性地提出饮食配伍，达到个体化营养的目的。

目 录

前言	(1)
第一章 局部基因网络概述	(1)
第一节 基因网络的研究概况	(1)
第二节 局部基因网络	(8)
第三节 基因多样性与个体化营养	(10)
第四节 复杂性疾病	(12)
第二章 局部基因网络的构建及其与各营养代谢耦联的 方法	(19)
第一节 局部基因网络的构建方法	(19)
第二节 局部基因网络与各营养代谢耦联的方法	(21)
第三章 乳腺癌局部基因网络分析	(22)
第一节 乳腺癌概况	(22)
第二节 乳腺癌局部基因网络的构建	(25)
第四章 乳腺癌局部基因网络与酒精代谢耦联关系	(45)
第一节 乳腺癌与酒精	(45)

第二节	乳腺癌局部基因网络与酒精代谢的耦联关系分析	(47)
第五章	乳腺癌局部基因网络与各营养代谢耦联关系	(62)
第一节	乳腺癌与各营养代谢	(62)
第二节	乳腺癌局部基因网络与各营养代谢的耦联关系分析	(63)
第六章	高血压局部基因网络与各营养代谢耦联关系	(78)
第一节	高血压与营养代谢	(78)
第二节	高血压局部基因网络与各营养代谢耦联关系分析	(79)
第七章	肥胖局部基因网络与各营养代谢的耦联关系研究	(100)
第一节	肥胖与营养代谢	(100)
第二节	肥胖局部基因网络与各营养代谢耦联关系分析	(101)



第一章 局部基因网络概述

第一节 基因网络的研究概况

一、基因网络的概述

20 世纪 60 年代基因网络的研究才刚刚开始, Rater 等用其描述了控制原核生物的分子基因系统组织的特点; Kauffman 等运用简单的逻辑规则研究了基因网络动力学。随着科学技术的发展, 实验数据迅速增加, 自然加速了基因网络理论的研究。迄今为止, 随机模型、布尔网络、逻辑方法、门限模型以及基于微分方程的方法等^[1~2]得到发展。

基因网络即基因调控网络, 通俗讲也是分子组合, 由蛋白质、基因、其他分子及共同执行某些细胞功能的分子之间关系组成的一个集合或组合^[3]。它主要对特定细胞或组织内的功能路径进行描述, 表达一些过程如基因调控、新陈代谢、传输和信号转换等。1953 年, Watson 和 Crick 发现了 DNA 的双螺旋结构, 开创了分子生物学的新纪元。这一发现, 对科学家从分子水平研究生命现象本质, 认识基因调控的作用具有重要的意义^[4]。然而基因网络是生物信息学的研究内容, 因此,



为了更好地理解基因网络的内涵和研究技术，我们完全有必要进一步了解分子生物学以及生物信息学的相关知识。

我们首先应对生物信息和基因之间的相互作用关系进行深入理解。

1. 生物信息学常常对基因网络的关系进行一些适当处理，如：简化和定量化等

通常的做法是，在实验数据和已有经验知识的基础上挖掘关系的信息，以此来建立网络模型，然后用模拟分析和实验来测试挖掘的各种关系，进而研究网络的特性，最终达到认识调控关系和相关机制的目的。

2. 认识复杂的分子调控过程，进行基因网络的生物信息学研究

利用建立的模型，旨在深入理解基因调控的时空机制，目的是识别和发现疾病治疗中潜在靶标的预测工具。生物系统的行为尽量通过不同的网络模型来描述，这样才更利于搞清具有协同效应的基因的作用机理及基因调控的复杂过程。

3. 搞清决定基因表达和功能的基本规则，许多基因操作的关键是要弄明白哪个基因何时进行表达的决定机制

我们将基因组中每个基因都当作互不相同的变量，它只有表达或不表达这两种状态，当然也可将其当作连续变量。现如今许多基因联合调控的研究事实上就是通过生物信息学的方法和手段来揭示基因的潜在调控机制。理解支配基因表达和功能潜在的规则是基因网络研究的最终目的之一。

4. 解开在基因表达过程中信息的传输规律

采取量化的方法对基因表达过程中的信息进行挖掘，进而



完成对基因网络的研究。通过基因网络模型的方法可以提取到基因联合表达中隐含的遗传信息。将其中的信息进行提取和分析，来了解基因的功能和理解基因网络，最终解释疾病的发病机制。

5. 在整体的架构下研究基因的功能

基因功能的研究一直是科学研究的一项重要任务，基因网络研究和功能基因组研究的进一步发展，为我们从整体出发进行研究和信息挖掘提供了保障，完全有可能发现基因的新功能。对于基因的功能，我们要分别从结构层次和网络层次来进行研究和认识。我们知道，基因是在错综复杂的网络系统中有条不紊地协同工作，因此，其中一定还有 DNA 序列没有揭开的新的功能。基因的功能通过基因表达来展示，但表达过程实际上是一个互相作用和调控的复杂过程。

在基因芯片技术诞生以前，生物学家们大多在研究单个基因及其表达，再深入一点，顶多研究几个基因之间的简单关系。但是，在基因测序技术特别是高密度 DNA 芯片和蛋白质质谱等技术的支持下，研究者们就能够在短时间内获得生物体基因表达的大量数据。这样就为研究多基因及其产物之间的相互作用关系，特别是为研究基因表达的时空机制提供了可能。因此，传统的研究方法——分子解析受到了系统地研究错综复杂的基因网络方法的挑战，特别是许多模式生物的全基因组测序工作的完成或即将完成，包括可通过微阵列技术获得一些模式生物的整个基因组的基因表达数据如酵母（yeast）等，所有这些都为从基因组水平研究基因网络创造了条件。

总而言之，一个基因的表达通常受其他相关基因的影响，



同时，这个基因反过来也影响着其他许多基因的表达，基因间的这种相互影响、相互制约的关系形成了复杂的基因表达调控网，更广一点来说，基因网络控制着几乎所有的细胞活动。生命是一个储存信息并对其进行加工的复杂系统，因此，要想搞清楚生命现象的内在规律仅凭孤立地研究单个基因及其表达通常是不行的。所以，科学家们不得不开始从复杂系统的角度研究基因网络，促使生物学家们以系统的观点来认识高度复杂的生命现象。

后基因组信息学研究的主要内容之一就是关于基因网络的研究，同时，通过数据采集、分析、建模、模拟和推断等手段，利用生物信息学的方法和技术研究复杂的网络关系，解释其作用机理同样也是目前生命科学的前沿领域。

二、基因网络的原理

一般情况下，一个基因的表达上调或下调会影响周边多个基因表达状态的变化，进一步导致和这几个基因相关的更多基因的表达模式的变化，说明基因的功能并不是独立存在的。基因间这种复杂的相互作用关系构成了一张错综复杂的立体网。如果我们仍像过去那样孤立地理解某个基因的功能已经远远不够了，我们必须站在更高的层次系统的理解基因间的相互关系，只有全面地了解不同个体基因变异、不同组织、不同时间、不同生命状态基因表达的差异信息，才能找出其中内在的规律。

受技术水平所限，科学家们以前一般按照基因及其产物是否对特定的表型起支配作用而进行分类，这样在面对由几个基



因及其产物共同作用的现象时就会遇到困难。随着相关学科研究的深入和相关技术的综合应用,我们需要从整个基因组规模、基因表达水平和蛋白质水平上来认识生物系统的结构和动态信息。就系统而言,大量的基因表达数据正说明基因网络这一信息处理的复杂系统是在分子水平上运转。所以,我们用带有反馈回路的基因网络来处理它。

基因网络理论的重要目的就是对某一物种或组织中全部基因的表达关系进行整体性研究。构建基因网络首先应该按照同步或反同步表达,以及表达强度的变化,系统地识别各基因的特点,然后再用聚类分析的方法将各基因归类,以此进一步分析相关的控制参数^[5]。构建一个合适的基因网络框架,与是否具备非常完整的基因表达数据并无必然的关系,而实际上,两者是相辅相成的,是一个相互作用的过程,如果实验数据不完善,完全可以适当引入假设^[6]。

就真核生物而言,它的全基因组极有可能是通过基因簇进行基因调节的,也就是说,基因组这一复杂的、紧密联系的控制系统往往是由基因簇组成的。通常这些基因簇存在多个调节位点,转录因子就是在这些调节位点上与之特异地结合。Davidson 等^[7]从基因组水平研究了海胆发育过程中转录的空间调控现象,这样就为后期从基因网络的角度进行整体分析奠定了基础。这些由具有相关功能的基因组成的基因簇在生理上存在的优势可能取决于信号分子的区域化和穿透能力,进而可以使各转录因子组成高度特异性的协同调节模块。



三、基因网络的研究进展

生命体系如此复杂的系统，对研究工作者来说的确是一个较大的挑战。要想真正揭开生命现象的复杂、整体之谜，就应该用整体和综合的理念与方法去研究，将传统的还原法与新的复杂系统理念进行有机结合，才能起到较好的效果。如此复杂系统的研究，无论从理念上还是从方法学上来说都是对传统的挑战，同时，还需要多学科交叉和方法学上的突破。我们知道，生物体中复杂体系的网络调控是多层次存在的，即便是同一层次也是由多组局域网连接而成。关于这方面的报道，国际上有为数不多的几篇相关的综述性论文发表，由此可以看出国内外关于基因网络的研究都还处在探索阶段。

分子生物学、计算机科学的快速发展无疑为基因网络的研究提供了很好的工具。由实验得到的表达数据不只是能用其分析基因表达的时空规律研究基因的功能，还能够用其分析基因间的相互制约关系研究基因转录调控网络。基因表达结果通常是细胞、组织、器官受遗传和环境影响的综合表现。我们知道，细胞的生化状态决定基因的转录，在转录过程中，通过转录因子作用于该基因的启动子区域来控制该基因的转录，同时这些转录因子本身又是其他基因的产物。一个基因完成转录、翻译形成功能基因产物后，它反过来还将改变细胞的生化状态，进而直接或间接的影响其他基因的表达，甚至还会影响基因自身的表达。若多基因的表达不断变化，那么，细胞的生化状态自然就会不断的发生变化。

总而言之，通常某一基因的表达会受其他基因的影响，与



此同时这个基因反过来又会影响其他基因的表达，基因间这种相互影响、相互制约的关系就形成了复杂的基因表达网络^[8]。从系统的角度来讲，一个细胞就是一个复杂的动力学系统，而这个系统中每一个基因又相当于系统的一个变量，同时，这些变量之间又相互影响。建立调控网络的数学模型是分析基因转录调控网络的目的所在，然后通过数学模型来分析基因间的相互作用关系。

生物信息学研究中的一个很重要的方向就是基因调控网络的分析和重建，这些研究内容已经引起了国外许多学者的高度重视，同时，其研究成果也被运用到了非常广泛的领域。生物信息学的研究在国内才刚刚起步，2003年，清华大学刚成立了生物信息研究院，主要从事基因组学的研究，最近才刚完成基因组的部分测序工作；近年来，其他部分高校也相继建立了生物信息学的研究机构，但大多还没有进行基因网络的研究，所以，基因网络研究在国内仅有为数不多的人略有认识，深入研究更无从谈起。如今，国外研究基因网络的模型主要有贝叶斯模型、线性组合模型、加权矩阵模型、微分方程模型、布尔网络模型、互信息关联网络模型、神经网络模型以及分子级的随机组分模型等。

随着生物信息学的迅速发展和广泛应用，越来越多的人意识到它的重要性，就像有些专家预测的一样：“21世纪将是生物的世纪”，所以，从事该方向的科研工作者也在日益增多，同时，对生物信息学的理解也在日益加深。我们完全相信，人类在不久的将来，通过基因表达调控网络能够解读越来越多的人类遗传密码。



第二节 局部基因网络

DNA 芯片技术的广泛发展，产生了许许多多与基因相关的原始数据，但就目前而言，科学家们还没有办法从这些数据中提取到满意的信息，其中主要是由于对特定疾病或表型的多基因功能性研究的例子很少造成的，另外就是可能生物学家没能力来解释清楚这些生物信息学数据，因为他们没有充足的相关资料。

人类细胞的基因数虽然大约只有 4 万个，但由这些基因而得到的多基因关系模式是海量的，所以，用目前可用的工具是很难分析具有上百个基因的基因网络。

我们研究的乳腺癌、肥胖和高血压等都具有复杂的表型，它们都有大量基因被报道说可能包含在它们的病理过程中。在本节的研究中，我们计划从小处开始并且试着用简单的方法来分析多基因关系。此方法被称为局部基因网络模型。

LGNP (Local Gene Network Pattern) 是基因网络的一种，LGNP 主要包括以下要点：

(1) 它能以线性的方式直观地反映基因与基因间的相互作用。

(2) 它包括许多潜在的基因间的相互作用。

(3) 在局部基因网络中可以直接显示出许多潜在的疾病致病基因，而且可以直接显示其与其他易感基因的相互作用。



(4) 根据疾病样本局部基因网络模式的表达状态, 它可以直接指出疾病人群的特定代谢模块, 由此可以阐释疾病特异性的营养缺陷和疾病特异性的药物开发。

(5) 细胞分子网络被限制在由有限的几个基因组成并且这些基因在功能上都有一定的相关性。这体现了“局部”的意义。

(6) 从网络的任一点一定有一条可以连接到其他一些点的路径。

(7) 每一个局部基因网络可以通过信号转导途径、代谢途径等一些细胞过程中有限的几条路径连接到其他的通路即所谓的模块性。

(8) 这些组成 LGNP 的基因可以被分成不同的组, 如: 早期基因、中期基因和晚期基因即所谓的时序性。因为他们来自不同时期的通路。所以, LGNP 可以在同一个网络内分析不同时期的基因在不同条件下不同的表达情况。

该模型主要的优点就是能以简代繁, 通过局部网络中有限的几个基因我们就可以对整个大的基因网络有所了解。创新之处在于: 通过局部基因网络与高通量数据如基因芯片数据的结合可以得到一批不同的局部基因网络模式, 这些模式与具体生物性状表型之间具有潜在的映射对应关系。性状表型包括我们感兴趣的营养缺陷或疾病。所以, LGNP 提供了一个创新性的手段研究多基因之间的关系, 提供了一个把个体基因型与生物性状包括营养代谢缺陷之间对应起来的简单而强大的工具。



第三节 基因多样性与个体化营养

人与人之间所携带的基因是有区别的，通常人的基因上约有 140 万至 200 万个单核苷酸多态性（Single Nucleotide Polymorphisms, SNPs）位点，而这些位点中大约有 6 万多处在外显子中^[9]。基因中单核苷酸多态性的存在直接导致不同的个体对相同营养素的反应是不同的，因此，食品中的有些营养成分对有些个体可能是有益的，而对另一些个体就不一定了，处于不同时期或阶段的同一个体可能对相同食物的同一营养素也会有不一样的反应。所以，营养学将来的研究将慢慢走向个体化、个性化^[10-11]。

研究人员应用蛋白质组学、营养基因组学等技术能够同时分析成千上万条基因及其表达产物。随着分子生物学技术的发展，在此基础上，运用基因芯片、蛋白质组学等技术探索营养素和基因间的相互作用是营养基因组学研究的主要内容。目前，营养学的研究逐渐走向个体化，研究个体基因结构和代谢类型，通过调整饮食使个体的营养状况处于最佳状态。进一步认识营养素，将基因表达、蛋白质产物与代谢组学相结合，研究出个体针对性更强的营养品。

营养基因组学在功能食品开发以及判定功能食品的个体效用性方面具有很强的功效，在于它能够解释生物活性膳食成分是否有效，也可以检测转基因食品的安全性，能够帮助公众消