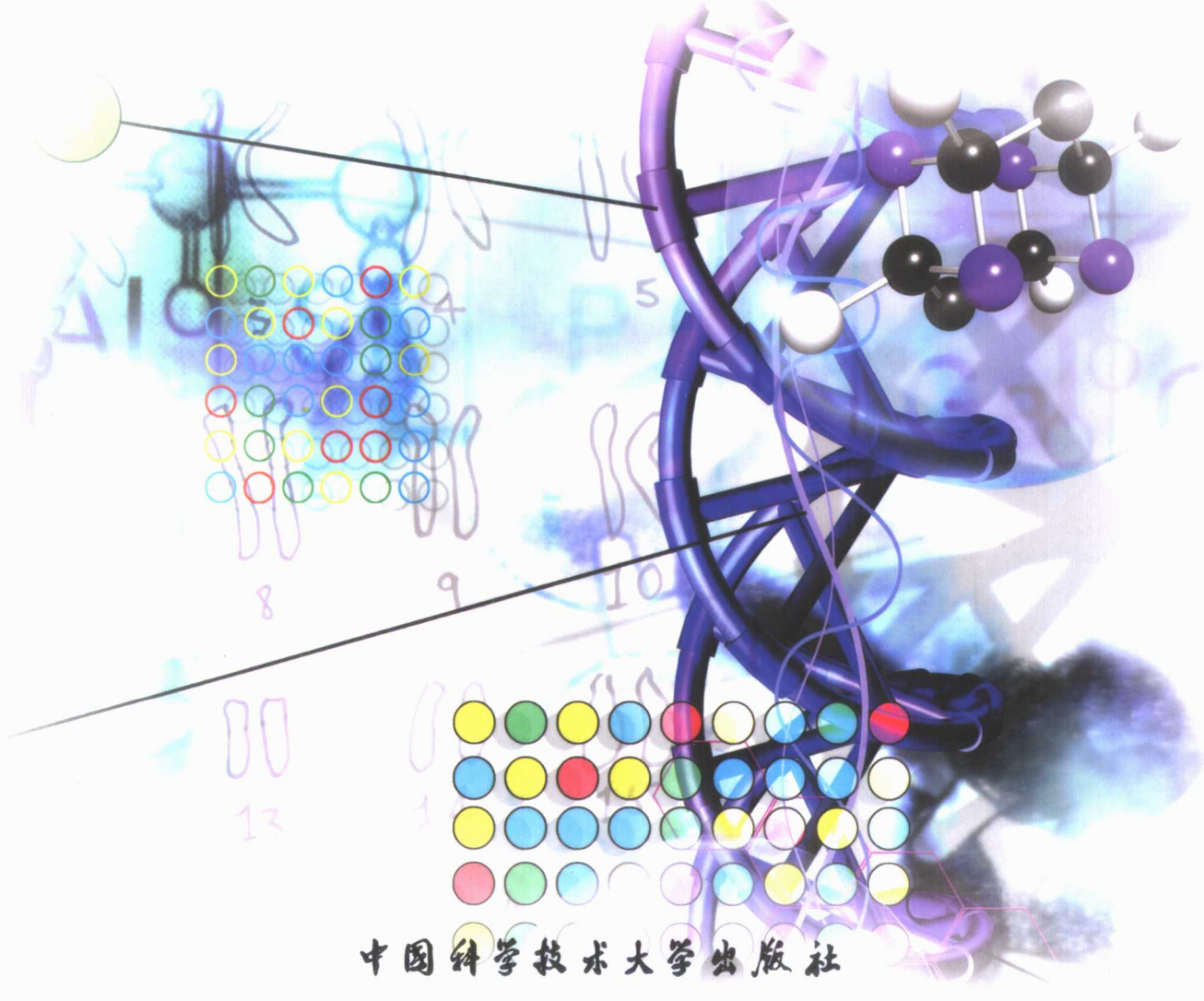
生物信息学若干前沿问题的探讨

Explorations of Frontier Problems in Bioinformatics

黄德双　张学工　田　捷　刘湘军　主编



中国科学技术大学出版社

生物信息学若干前沿问题的探讨

Explorations of Frontier Problems in Bioinformatics

——中国科协第 81 次青年科学家论坛论文集

中国科协“生物信息学与进化计算”青年科学家论坛

2003 年 11 月 28-29 日，北京中国科技会堂

黄德双　张学工　田　捷　刘湘军　主编

中国科学技术大学出版社

2004 · 合肥

内 容 简 介

中国科协“生物信息学与进化计算”(Bioinformatics and Evolutionary Computation)青年科学家论坛(简称:BEC'2003)已于2003年11月28～29日在北京中国科技会堂顺利召开。本次论坛的主题为“生物信息学若干前沿问题的探讨”。32位来自全国各地的生物、医学、信息、计算机、数学、物理学科的青年科学家参加了这次活动。论坛邀请了清华大学李衍达院士、中国科学院生物物理研究所陈润生研究员、内蒙古大学罗辽复教授和北京华大基因研究中心的于军研究员在论坛上作了精彩的特邀报告。此外还有24位代表在会上介绍了自己的工作和心得体会,并做了广泛的学术交流。论坛结束后,我们向各位代表发出征集出版本论文集的邀请,最后经选择收录了陈润生研究员和罗辽复教授的大会报告讲话稿,收集了21位代表的论文,并以“生物信息学若干前沿问题的探讨”的名字出版这本册子。特别令我们感到高兴的是,清华大学李衍达院士还在百忙中为本次论坛论文集的出版写了序。

期望本论文集的出版能对我国生物信息学的研究和发展起到一定的推动作用。

图书在版编目(CIP)数据

生物信息学若干前沿问题的探讨/黄德双,张学工,田捷,刘湘军主编. 合肥:中国科学技术大学出版社,2004.11

ISBN 7-312-01749-5

Ⅰ.生… Ⅱ.①黄… ②张… ③田… ④刘… Ⅲ.生物信息论—文集 Ⅳ.Q811.4-53

中国版本图书馆CIP数据核字(2004)第111781号

中国科学技术大学出版社出版发行
(安徽省合肥市金寨路96号,230026)
中国科学技术大学印刷厂印刷
全国新华书店经销

开本:880mm×1230mm 1/16 印张:15 字数:640千
2004年11月第1版 2004年11月第1次印刷
印数:1—350册
ISBN 7-312-01749-5/Q·43 定价:88.00元

中国科协学会学术部马阳部长在讲话

中科院院士清华大学李衍达教授在讲话

中国科协学会学术部李慧政局长在讲话

中国电子学会副秘书长李志武在论坛上讲话

中科院院士清华大学李衍达教授在做大会报告

中科院生物物理所陈润生研究员在做大会报告

内蒙古大学罗辽复教授在做大会报告

北京华大基因研究中心于军研究员在做大会报告

执行主席黄德双研究员在发言

执行主席刘湘军教授在发言

执行主席张学工教授

执行主席田捷研究员

会场一角（1）

会场一角（2）

会场一角（3）

部分与会代表合影

序　言

目前生物学和信息学结合起来的生物信息学，是自生物学、信息学之后，又一门热门的科学。自从1990年美国启动人类基因组计划（又被称为生命科学史上的“登月”计划）以来，经过美、英、日、法、德和中国科学家的艰苦努力，已于2000年6月26号完成人类基因组工作框架图。它预示着完成人类基因组计划最后目标已经指日可待。然而，随着基因测序工作的深入，人们得到的DNA序列数据的增长将十分惊人，如目前仅登录在美国GenBank数据库中的DNA序列总量已超过70亿碱基对。生物学数据的积累并不仅仅表现在DNA序列方面，与其同步的还有蛋白质的一级结构，即氨基酸序列的增长。此外，迄今为止，已有一万多种蛋白质的空间结构以不同的分辨率被测定。基于cDNA序列测序所建立起来的EST数据库其纪录已达数百万条，在这些数据基础上派生、整理出来的数据库已达500余个，这一切构成了一个生物学数据的海洋。这种数据的急速和海量积累，在人类的科学研究历史中是空前的。

然而，数据并不等于信息和知识，但却是信息和知识的源泉，问题的关键在于如何从中挖掘出知识。与正在以指数方式增长的生物学数据相比，人类的相关知识的增长却十分缓慢，这就构成了一个极大的矛盾。这个矛盾就催生了一门新兴的交叉科学，这就是生物信息学。生物信息学包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具，来阐明和理解大量数据所包含的生物学意义。

生物信息学的发展必将为生物、医学和信息技术的发展带来新的契机和新的挑战。生物信息学涉猎广泛，在基因组测序、基因组分析、基因识别、表达图谱分析、蛋白质结构预测、医学图像分析等方面做出了重要贡献。在这些领域的生物信息学研究虽然经历了一段时间的发展并取得了很大成就，目前仍然面临各种各样的挑战，也不断出现新的问题。随着生命科学的发展，特别是功能基因组研究的突飞猛进，生物信息学研究中还将不断涌现新的研究热点。

2003年11月末在北京中国科技会堂成功地举行了中国科协“生物信息学与进化计算”第81次青年科学家论坛。这次论坛汇聚了来自全国各地的生物、医学、信息、计算机、数学、物理学科的部分优秀青年学者，交流在生物信息学领域研究的心得和体会，共同探讨生物信息学的前沿问题，取得了预期的效果。这本论文集基本上反映了与会代表对我国生物信息学研究现状、发展方向和存在的问题的综述和展望，相信它的出版一定能对我国生物信息学的研究和发展起到推动作用。

李衍达

2004年10月

中国科协“生物信息学与进化计算”青年科学家论坛成功召开

2003年11月28－29日，中国科协“生物信息学与进化计算”第81次青年科学家论坛在北京中国科技会堂成功召开。这次论坛是中国科协举办的一次多学科交叉的盛会，旨在促进国内青年科学家在这一全新领域内的相互交流，促进该学科的成长与发展。中国科协和国家自然科学基金委的有关领导出席了论坛，并做了热情洋溢的讲话。

本次论坛的执行主席由中科院合肥智能机械研究所黄德双研究员、清华大学张学工教授、刘湘军教授、中科院自动化所田捷研究员担任，论坛的主题是“生物信息学若干前沿问题的探讨”。

生物信息学是一门新兴的边缘交叉学科，是近年来国际上的研究热点。生物信息学是利用信息技术来理解生物问题的一种手段，是探索生命奥妙，扩展生物医学行为数据使用的一种方法，是通过计算方法将生物信息转化为知识的工具。伴随着人类基因组测序的完成，大规模的生物序列数据也呈现爆炸性的增长。如果利用自动化程度高，既省时又省力且精度高的计算机技术来分析这些生物序列数据，则能够挖掘出许多不为人知的隐含知识，能够为人类了解产生疾病的根源等问题做出贡献，因此生物信息学的发展具有重要意义。生物信息学的发展必将为生物、医学和信息技术的发展带来新的契机和新的挑战。

这次青年科学家论坛有30多位工作在生物信息学科研第一线的青年科技工作者参加，并邀请了清华大学李衍达院士、中科院生物物理研究所陈润生研究员、内蒙古大学罗辽复教授和北京华大基因研究中心的于军研究员在论坛上作了精彩的特邀报告。他们一致认为系统生物学、非编码区功能研究、基因调控和相互作用网络等是当前生物信息学研究的热点问题。

此外，本次论坛执行主席黄德双研究员、张学工教授、刘湘军教授、田捷研究员等青年科学家在论坛上还就生物信息学中存在的问题、各学科的交叉发展情况和国内外的发展动态等问题，和与会代表进行了热烈的讨论。这是国内首次以“生物信息学与进化计算”为主题的一次多学科交叉的青年科学家论坛。相信通过这次青年科学家论坛的召开，国内的青年科技工作者能够对生物信息学的发展方向和存在的问题有更深入的了解，从而促进我国在这一前沿领域的发展。

最后，本次论坛执行主席黄德双研究员、刘湘军教授对本次论坛做了较全面的总结，中国科协学会学术部部长马阳研究员对本次论坛的成功召开给予高度的评价。论坛在非常愉快的合影留念后结束。

目　录

非编码基因和系统生物学 …… 陈润生　1
生物信息学接受后基因组时代的挑战 …… 罗辽复　4
酵母基因中转录正调控内含子的序列特征 …… 张　静　10
关于生物序列的一种位置约束比较方法研究 …… 李玉鉴　18
基因选择算法研究 …… 封举富等　26
DNA 序列分析的现代信号处理方法初探 …… 饶妮妮等　36
从基因可变剪接理解真核生物基因功能的复杂性与多样性 …… 张成岗等　43
利用序列统计特征分析基因组序列 …… 孙　啸等　51
小动物活体成像系统与生物医学光子学 …… 田　捷等　57
针刺镇痛的脑功能磁共振成像初步研究 …… 艾　林等　68
免疫信息学：生物信息学新领域 …… 黄　健等　76
基因调控元件的计算机识别和基因调控网络构建 …… 罗静初　李伍举等　92
基于神经网络技术的蛋白质二级结构预测 …… 黄德双　张广政　109
从病毒细胞遗传进化的生物循环选育到新基因识别与验证 …… 张德礼等　117
Carcinogenesis of Renal Cell Lines in Nude Mice in a Physiological Context
Clarifying the Origin of Malignant Rhabdoid Tumor (MRT) …… 张德礼等　150
关于基因组生物信息学研究的展望 …… 王　俊　于　军等　172
中医药现代研究与生物信息学 …… 李　梢　179
植物基因家族的进化 …… 顾红雅等　189
基因组是通过基因组语言编写生命过程的一个特殊软件 …… 张治洲　201
DNA "分子手术" …… 胡　钧等　208
Exploiting the interactions of single DNA molecules with mica surface …… 方海平等　215
Recursive Sample Classification and Gene Selection Based on SVM:
Method and Software Description …… 张学工等　220
真核生物顺式调控模块的识别 …… 刘　蓉　226

非编码基因和系统生物学

陈润生

中国科学院生物物理研究所，北京，100101
crs@sun5.ibp.ac.cn

长期以来令人惊异与困惑的是：生命并不是一群分子的堆集，它是高度有组织的。生物分子组成细胞，细胞构成组织，组织形成器官并进而构成系统。细胞核与细胞质有作用；细胞与细胞间有联系；组织与组织间有分工；器官与器官间有协同。因此一个正常生存的生物个体是极端有序的，是多层次的，是动态的。他是经亿万年的进化才由无序到有序，由简单到复杂的。不仅生物个体如此，生物群体也是高度有序的整体，一个生态系统有互利、有竞争。因此，生物体的复杂性并不仅仅表现在DNA信息结构的复杂性上，还表现在这些信息的实施与运作的规律上。那么，生物体的复杂结构和功能是如何产生与维持的呢？上世纪六、七十年代开展的非平衡热力学理论就指出平衡态是无序的，而非平衡态才可能是有序的。正常的生物体是“活”的，他能生长、发育、繁殖和新陈代谢，他是一个不断地与外环境进行物质和能量交换的开放系统，生物体是远离热力学平衡的，生物体中大量的过程是不可逆的。因而生物才能生存、能进化，是有序的。这说明：非平衡是有序的起源，或者说是信息的起源。因此，从根本上来说，生命是由于复杂的物质相互作用产生的。当简单生命出现以后，生命的进化也是由于生物体与环境的相互作用形成的。相互作用是生命产生与演化的基础，因此，相互作用也必然贯穿于生命活动的每一个环节。

自上世纪五十年代Watson和Crick提出著名的DNA双螺旋结构模型以来，人们对生命本质的认识一下子进入到分子水平。接着三联体密码和密码表的发现；基因作为DNA分子中编码蛋白质区域的确定；遗传信息流从DNA→RNA→蛋白质这样的“中心法则”的提出；基因表达调控的操纵子模型的构建以及DNA半保守复制被证实，构成了一幅遗传信息存储、表达、调节、复制到实现复杂功能的完美图画。使得人们第一次在分子水平上了解了遗传信息的传递、调节以及这些分子间的相互作用。

自从上世纪八十年代末人类基因组计划一出现，破译人类遗传密码就成为当代人们共同关心的科学问题。公共数据库中DNA碱基数目呈指数增加，大约每14个月翻一番。它超过了电脑芯片计算能力的增长速度。到2003年初，DNA碱基数目已超过170亿。到2003年底，已公布的全基因组序列有113套，其中古细菌16套，真细菌82套，真核生物15套，包括：酵母、线虫、果蝇、拟南芥、水稻、小鼠等的完整基因组。正在测序中的有原核生物344套，真核生物235套(见：http://igweb.integratedgenomics.com/GOLD/)。2001年2月在Nature和Science杂志上同时发表了由国际人类基因组组织(HUGO)和Celera公司分别完成的人类基因组序列及其初步的分析，给我们展示了关于人类基因组的一系列较以往更为细致、更为精确的信息。这些序列中含3—4万个编码蛋白质的基因。基因组的1.1%为外显子，24%为内含子，75%为间隔序列。目前，数据库中还收集了500多万个代表着人类基因表达小片段的EST(Expressed Sequence Tags)数据，它大约覆盖了人类基因的95%，冗余度已超过10。近年来增长最快的数据是人类的SNP

（single-nucleotide polymorphisms），它代表着不同人种以及正常人和某些病人基因组中碱基的差异。2003 年已有 300 多万个人类非冗余 SNP 位点。这表明人的基因组中平均每 1000 个碱基就有 1 个碱基差异。但在已知 SNP 中，仅有不到 1%的 SNP 造成蛋白的变化。同时，由于生物芯片、二维凝胶电泳和测序质谱技术的高速发展和广泛应用，功能基因组和蛋白质组的数据也已大量涌现。这些海量数据为更深入地在分子层次分析生命活动的规律提供了机会。

通过分析海量数据，科学家们发现：DNA 上编码蛋白质的区域（也就是基因），只占人类基因组的 3%，其余 97%左右的 DNA 序列仍不大清楚其功能，国际上科学家们习惯地把这部分 DNA 统称为“非编码 DNA”或“Junk” DNA。通过对完整基因组的比较发现，低等的生物，象病毒、细菌等只有少量的“Junk”DNA，而高等的动、植物则含有大量“Junk”DNA，它们甚至占据着基因组的大部分。这就是说，伴随着生物从简单到复杂、从低级到高级、从信息少到信息多，非编码 DNA 不断增加。它意味着“Junk”DNA 可能蕴涵着生物体复杂性的信息并可能参与生物大分子的各种相互作用。近年来大量的新实验结果表明非编码 DNA 是可以表达的，其表达产物是许多对生命过程富有活力的信息载体。小 RNA(small RNA)的研究就是最突出的例子。“Science”周刊 2001、2002 连续两年将“small RNA”评选为该年度全球科学十大进展，而 2002 年更将其作为进展的第一位。这表明非编码 DNA 的研究已引起国际上的广泛关注。现在科学家们正系统地从真菌到植物、从无脊椎动物到哺乳动物，甚至也从低等的原核生物中寻找小 RNA 基因，并试图确定成百种小 RNA 各自的功能；确定哪些物种含有哪类小 RNA 以及它们在该物种中的行为是什么？小 RNA 的出现重新唤起了科学家们对“RNA 世界”的重视及对“生命起源于 RNA 分子”这一命题的兴趣。有的科学家认为成千上万非编码蛋白质的 RNA 分子组成了巨大的分子网络，调节着细胞中的生命活动。它们与蛋白质-蛋白质相互作用网络相对应，好比宇宙学中的暗物质与亮物质。因此，如果说 DNA 的编码区带有组成人体具体材料(蛋白质，结构 RNA 等)的信息,那么大量的非编码区应当包含把这些材料按照特定的时间、空间安置，以组成完整个体的四维调控信息。这种信息应有动态特征，并能产生时间标度。总之，这些新的研究进展告诉我们即使在分子层次上生物体的相互作用由于非编码基因的出现而变得更为复杂、更为丰富了。

通过人类基因组计划的实施，人们得到了各种生物的基因图谱，它给了我们前所没有的大量信息。但仅有这类静态信息仍不能说明生物（生命）是怎么工作的，它又是如何活起来的。为此，我们必须了解基因是如何按照特定的时间、空间进行表达的，表达量有多少。不了解伴随着生物的生长发育，基因表达状况的变更，也就无法确切地说明生命的过程。同时我们还需要了解基因网络，需要了解生物分子反应的调控途径和代谢途径等。由于生物体的层次性和整体性,知道了这样的一个个系统、一个个调控单元还不够，还要把所有的单元之间的关系耦联起来，整合在一起。也就是要把发生在基因表达层次、蛋白质折叠层次、酶催化层次、特定蛋白质在细胞里的定位层次、生物大分子代谢层次等等的事件整合起来。这样才能模拟从细胞、组织、器官、系统到整体的生物系统的行为，并可用来预测如果这个系统一旦受到了刺激和外界的干扰将如何演变。近年来，系统生物学就是基于这样的观念而产生的。系统生物学的观点认为：通过不同层次关联建立起来的复杂系统，并不是简单系统的叠加。这个复杂系统会出现一些突现性行为、突现性规律，就是出现一个单独系统所不能反映的新行为。应该说生物体的复杂结构和功能不仅仅是由基因和基因组中大量的非编码信息决定的，还是由生物体各个层次的复杂、动态相互作用决定的。

系统生物学，从理论来讲，是从我们传统的序列、结构到功能这样一个思维方式，变成从相互作用、构建网络到说明功能这样一个完全不同的方式。过去，都认为一个基因表达一个蛋白，一个蛋白有一个结构，一个结构完成一个功能。现在越来越多的事实说明，一个基因的单独表达往往不能主宰一个生物学事件的发生，一个事件的发生是一堆基因的同时表达，是一堆蛋白质的协同作用。因此实际上真正表现生物

学功能的，在研究方式和思维方式上应当代之以相互作用的概念。那么相互作用之间就构造成一个网络，所以将来也许真正要说明生物学功能的并不是去追求某一个单独基因的一个蛋白质，而是要考虑一组相互作用的网络，这就是系统生物学所带来的思维上的变化。这就是所谓的研究思路的变化。系统生物学的研究思路实际上是着重多信息融合过程的构建。所谓系统生物学并不是去代替其他科学家，要把基因水平上的东西搞清楚，要把蛋白质水平上的东西搞清楚。他是做这些分子生物学家或遗传学家不做的事情，他是要把他们的东西拿来，去找到他们之间的联系，要把他们融合起来，所以做两个系统或多个系统之间的那份工作。因此这个工作主要是找层次与层次之间、系统与系统之间的联系。什么是融合？融合就是找发生在两个系统、两个层次之间的关联。系统生物学的研究思路就是去研究两个层次、两个系统或多个层次之间的耦联。耦联是什么呢？耦联就是相互之间的作用。系统生物学就是找在不同系统、不同层次之间的关联，这个关联主要表现为某些特殊的作用。要做这件事情，我们要把转录水平、相互作用水平、蛋白质水平、基因水平，所有这些东西整合在一起。怎么整合？就是找它们之间的联系，找它们之间的相互作用关系，这就是系统生物学的核心的研究思路。

最后相互作用还存在于生物群体之中。对人类来说还可能涉及思维、意识、学习、记忆等复杂问题。对于这类生命活动高级形式的了解还需要更长的时间、更多的努力。

作者简介：陈润生，男，中国科学院生物物理研究所研究员，博士生导师。1964年毕业于中国科学技术大学生物物理系，1985-1987年获德国洪堡奖学金在纽伦堡大学理论及物理化学研究所作访问学者，中国生物物理学会常务理事，中国物理学会理事，国际人类基因组组织（HUGO）委员，国际数据库组织（CODATA）生物大分子专业组委员，国际纯粹及应用物理学会（IUPAP）生物信息学专业委员会委员，中国生物物理学会秘书长、副理事长。陈润生研究员由于在基因组信息学和蛋白质三维结构模拟领域的贡献，1996年10月3日在日本筑波召开的第十五届CODATA国际学术大会上被授予“小谷正雄”奖。

生物信息学接受后基因组时代的挑战

罗辽复

内蒙古大学理工学院物理系，呼和浩特，010021

lfluo@mail.imu.edu.cn

1. 生物信息学的兴起和生命科学的理性化

上世纪 90 年代中期以来，从细菌到人类，几十个物种的核酸和蛋白质序列数据迅猛增长。1965 至 1978 的 13 年中共测定发表了 12000 个核苷酸， 1982 年底达 100 万核苷酸， 1990 年底达 5000 万核苷酸。此后速度愈来愈快，2001 年春，含 30 亿碱基对的人类基因组工作草图提前完成。截至 2002 年底，全世界 GenBank 文库中的总碱基数已达 300 亿个，基因数据量每 6 至 8 个月翻一番。这样高的几何级数增长速度，在科学史上是前所未有的。反映个体和病理差异的单核苷多态性(SNP)数据也从 1999 年的 2 万个发展到现在的上千万。除了 DNA 测序外，蛋白质结构的数据也在迅速增长，现在每个月已至少能测出 160 种以上的蛋白质三维结构。一些高通量实验技术，如 DNA 微阵列(基因芯片)，能对基因组范围的上万个基因表达同时测定。面对着海量的生物学数据，传统的生物学研究方法已经完全不能适应。于是生物信息学应运而生。其直接目的是解决生物信息的获取、处理、存储、联网、浏览等问题。更深层次的目的则是对如此大量数据的分析和解释，以及数据背后隐藏的生物学规律的探寻，即数据挖掘。

生物信息学是新生的交叉学科，基因组信息学则是生物信息学的核心。10 年来，在全世界的各个国度，它吸引了数以千计的计算机科学家、数学家、物理学家、生物化学家和遗传学家的参与。由于参与学者的不同学科背景，对这门科学的内容和方法、重点和目标，以及它在生命科学中的地位就有不同的理解。一般说来，每一门成熟的学科都有相对稳定的科学共同体和使用较为一致的价值标准。 而对于新生的生物信息学，就缺少此方面的共识。 于是，随着升温到一定阶段，就出现了一些疑问：生物信息学能解决生命科学的基本问题吗？有人提出系统生物学的概念，认为生命是一个整体，必须用维纳系统论、控制论的观点和方法才能解决问题[1]。有人希望把 Computational Biology (计算生物学) 和 Systems Biology (系统生物学) 综合起来[2]。近两三年英国 Nature 和美国 Science 上 “连篇累牍”地发表关于蛋白质网络和基因调节网络的论文，从一个侧面反映了部分“主流派”生物学家对这个问题的倾向性和对系统生物学的情有独钟。

为了更深入了解生物信息学或基因组信息学在生命科学中的地位，让我们对近代自然科学发展史作一个简短回顾。近代科学是从伽里略——牛顿开始的，形成了一定的规范（此处用规范一词系借用科学史家库恩的术语）。这个规范的要点有两个，第一是实证性，第二是理性[3]。伽里略继承了文艺复兴的先进思想：“我们的一切知识全都来自我们的感觉能力”，“经验是一切可靠知识的母亲，那些不是从经验里产生，

不接受经验签定的学问，那些无论在开头，中间或末尾都不通过任何感官的学向是虚妄无实、充满谬误的”。唯有充分重视实验，才能摆脱中世纪经院哲学的桎梏，这是伽里略成为近代科学之父的原因。然而，由于仪器精度的限制和直接感官的局限性，由于对各种复杂因素难于全面分析，常常容易被表面现象蒙蔽。所以实验必须与推理相结合。其实，惯性定律的发现就是用了实验与推理结合的“理想实验”的方法。因此，伽里略——牛顿开启的近代科学传统的另一个要点就是理性，最好的最严密的推理工具是数学。伽里略十分重视数学，他说：“自然之书是用数学语言写成的，没有数学，一个人只能在黑暗的迷宫里徘徊”。牛顿“自然哲学的数学原理”一书的写法完全照抄欧几里德的“几何原本”。定义、公理、定理、推论……。试问，没有数学的精密的推演，如何能证明天体运动和地面上抛物是由于同一种力——万有引力呢？实证性和理性的结合，便形成了四百年来的近代科学规范。这种结合在物理学中，特别是二十世纪的物理学中表现得最为完美和富有成果。 相对论和量子论的发现就是这种结合的典范。依靠数学的帮助，它们都有很多推论。由于数学链条之长和复杂，结论和出发点的关系往往已是靠直觉无法觉察的了。 通过无数推论和实验的非常严格的比较来证实逻辑出发点的正确性。 但是，实证性和理性的结合在其它学科中还不像近代物理学中那样完满。上一世纪初，著名的原子物理学家卢瑟福不无骄傲地说，除了物理学外，其它科学只不过是集邮。一百年前，这话可能是事实，但是实证性和理性结合的规范正以战无不胜的巨大力量征服了和正在征服着自然科学的其他部门。例如，用量子力学理论和方法解决化学问题就是一个例子。理性化的规范也正在向生命科学渗透，传统的生物学都是实验的，“正在建立的研究新模式是：由于全部基因将被知晓，存储在电子数据库中，生物学研究的出发点将是理论的。一个科学家将从理论假说出发，然后转向实验，去追随和检验这些假说”。[4] 因此，近代科学规范向生命科学渗透，把传统的实验方法和定量的逻辑的理性的方法相结合，从而揭示和了解生命规律，把它从一门实验科学提高到综合的理性的水平上来，这是科学历史发展的必然。

事实上，从上世纪中叶分子生物学诞生以来，生物学理性化的努力就没有中断过。例如，60年代Pullman夫妇建立的以研究核酸中电子运动为主要对象的量子生物学[5]，70年代Prigogine的耗散结构理论和Kauffman，Wolpert等人创办的理论生物学杂志，80年代以Murray为代表的以模拟个体发育中形态建成问题为中心的数学生物学[6]，90年代Waterman等人提出的以研究分子序列为中心的计算生物学[7]，等等。一波又一波，他们的理性化的努力方向和生物信息学是相同的。由此可见， 生物信息学的诞生是科学发展历史的必然， 是生物学理性化的一个组成部分和一个阶段。和前几次的理性化相比不同的是，90年代以后基因组海量数据的出现使得传统生物学无所措手足，所以目前的生物信息学这一波显得格外有影响力、格外轰动。至于目前关于系统生物学的思路，其实也早已有之，可看成是生物学理性化的进一步努力。

人们都同意，生物信息学包含几个关键词：数据、计算和知识扩展。作为生物学理性化的一个阶段和组成部分，生物信息学研究应加强对科学规律的探索，给传统的实验生物学以更多的理性成分。这样做也能加深生物信息学的理论深度，提高对它的预测能力。

二十世纪物理学是自然科学的领头学科，人们推测二十一世纪的领头学科将是生物学。

生命科学中还有太多的基本问题有待解决，而且这门科学和人类的生活和命运又是如此紧密地相连，它会在自然科学中具有领头地位，是并不奇怪的。但是如果生物学真的成为自然科学的领头学科，那么，它必定不是传统的生物学，而是渗透了近代科学规范的新生物学，是理性化的生物学。

一个更加基本的问题是：生物系统有没有独特的基本规律？有！因为生命不能还原为无生命世界。但生命系统独特的规律决不会是非物理的。它必须和现有的物理规律不相违[8]。这种独特规律的探索必将

有助于生物学的理性化，也必将丰富人们对复杂系统物理学规律的了解。

依据何种线索去发现这个规律？过去的自然科学基本上聚焦于物质和能量。这是自然界的两个基本范畴。不同于物质和能量，与其相平行的“信息”是自然界的第三个基本范畴。 生命系统的特征就在于它包含的大量信息。DNA 分子从物质上来讲，无非是碳氢氧氮等元素组成，从能量来讲，也只有微小的一点，都没有什么特别。但 DNA 含有大量信息，这是生命活动有序化的根源所在。更奇妙的是：这种信息是通过亿万年的自然选择，在无生命的自然界中在随机的背景下形成的，是大量的偶然性凝练而成的。根据申农的定义，信息是通过对事物随机选取的可能性来度量的。大量信息意味着大量偶然性。对于偶然性，物理学习惯的处理是进行统计研究。但自然界教给我们另一种处理方法，就是对每一次的偶然性进行随机确定，通过自然选择找出一条最佳或较佳的路线(或序列)。这个被保留下来的序列反映了大量偶然性背景中形成的生命之序。

我们曾经提出，“密码—序列—结构—功能”是一条可能的生物学理性化路线[9]。密码是随机序列(英语 stochastic 有别于 random)具有信息内容的基础，撇开它的形成问题不管，这条路线的三个结点是序列、结构和功能。解决如何从序列到结构，以及如何从序列结构到功能，这是实现生物学理性化的两个关键问题。最近人们认识到生物功能的实现往往依赖于一个网络，例如蛋白质相互作用网络、遗传调节网络和某一疾病相联系的多基因网络等。上述路线可更明确地表示为“序列—结构—功能网络”。

下面，依据这一条路线介绍一下我们组的近期工作，谈一点看法。

2. 遗传密码的进化和逻辑

遗传密码的起源和进化一直有两派观点争鸣。一是立体化学理论（Woese，1966），二是冻结偶然性理论（Crick，1968）。后来 Wong 提出氨基酸和密码字典的协同进化理论。我们近年的工作说明以上三种观点各反映了一部分实际情况。另一方面，近期一些工作试图从密码对称性的角度研究它的进化，例如提出 Sp(6)群对称性。我们考虑，密码的连续对称性要求也许太高了，如果研究分立对称性，选用 S_4 群是最合适的。S_4 包含了两个 4 阶子群，一是循环群 Z_4，另一为 Klein 4 群 V_4。看来 V_4 是更好的侯选者。S_4 对称性破缺为 V_4 对称性后，并没有到此终止，V_4 对称性还要进一步破缺。破缺 V_4 对称性可用阴阳对偶性的概念表示出来[3]。

遗传密码的极高普适性和内在单纯性说明了密码表的构成遵循着一定的逻辑。我们从进化稳定性的角度研究这种逻辑，说明现有密码表的结构，解释如何由密码子编码氨基酸的编码规则。假设现有密码是密码系统长期历史发展的产物，这种发展总趋势是朝着稳定化方向—突变危险性参数（简称 MD）极小化的方向进化。据此可以确定现有密码表上各个简并密码子多重态的分布，证明现有密码的简并规则是满足 MD 极小化原则的。遗传密码表的总体突变危险性依赖于两个因素，一是密码表中包含的一对对不处于同一简并多重态（氨基酸）的密码子的遗传距离，即两个密码子是否可以通过单碱基突变相联系，以及通过何种突变相联系？二是这一对密码子所编码的氨基酸相似性或距离。最近（Luo and Li，BioSystems 2002[10]；Origins of Life 2002[11]），我们求出了极小表，发现和标准表颇为接近，都有相同的亲水-疏水畴。但是，为什么大自然选择了标准密码表而没有选择极小表？我们进一步证明了标准密码表是在某些和密码初始形成时的状态有关的约束条件下总体突变危险性极小化的结果。突变危险性极小是进化稳定性的数学表示，而约束条件则表现了进化过程中的冻结偶然性。另外，1979 年后发现了普适密码的一些反常。突变危险性理论可对密码反常的稳定性作出估计。因此，这个理论原则上可以用来研究遗传密码的改造等具有广

阔应用前景的问题，为基因工程开辟新的途径。因为突变产生的新型氨基酸(如磷丝氨酸、硒半胱氨酸、4氟色氨酸等)可能改变密码，弄清密码的逻辑将有可能从理论上对突变进行预测和设计，从而为人工设计制造新功能蛋白质开辟广阔的前景。

3. 基因预测和基因组分析

我们研究了以酵母为典型的简单生物基因组中ORF（开阅读框）组织的规律性。其中包括：双链三种框架的6种ORF排列的随机性；ORF从终止密码子下游第一ATG起始的法则；短ORF起源的随机性；以及识别ORF的IHI（碱基在3个密码子位点上分布的不均匀性）规则，由此简单预测规则可得95%以上的准确度。(Luo，Li，Zhang，Comparative and Functional Genomics，2003 [12])。对于真核，基因识别的一个关键性难题是内含子剪切位点的预测。对此我们采用多样性指标，结合使用二次判别法，提出IDQD（多样性增量二次判别法）方法，从线虫，拟南芥，果蝇到人，预测准确度都达90%－95%以上，优于国际上其他软件。我们期望，由此出发，对于解决和功能密切相关的可变剪切（一般可占到剪切位点的5%）问题，能获得一些较好的结果。(Zhang，Luo，Nucleic Acids Research，2003 [13])

“不从进化看问题，生物学便是不可理解的”。我们研究了基因组的进化，考虑了随机突变，自然选择和片段重复三种力的作用（Luo et al，Physical Rev E，1998[14])。发现在进化早期，片段重复对于形成一个足够长的基因组可能具有特殊的意义（Hsieh et al，Physical Rev Letters，2003[15])。但是，基因组的信息是如何积累起来的？这个问题仍然是对生物信息学的困难挑战。

另外两个富有挑战性的问题，一是：如何对基因启动子(promoter)和转录因子结合位点(TFBS)进行有效的识别，以及对蛋白质－DNA调节码的探索。二是：如何去发现各种类型的非编码RNA基因（包括miRNA，siRNA等)，对它们正确地分类，有效地确认和预测。前者关系到基因转录调节的基本规律的探索和真核基因表达调控网络的建立，后者涉及真核基因组中大量（95%以上）迄今仍然是”黑盒子”的非编码序列生物功能的了解。

4. 蛋白质折叠和结构预测

蛋白质折叠分为离体的变性蛋白再折叠和体内新生肽链的折叠两大类。Anfinsen从离体实验中提出了蛋白质折叠的热力学假设，认为蛋白质天然构象是Gibbs自由能取全局极小值的状态。新生肽链的折叠更加复杂，可能是co-translational的，要有酶和分子伴侣等辅助蛋白的参与，折叠速率也比离体蛋白折叠快得多。我们以量子构象动力学为研究蛋白质折叠的理论基础，导出了折叠的基元过程－构象跃迁速率的时标为毫秒量级，并证明了合作跃迁的速率与模数的3/2次方成正比，从理论上说明了蛋白质规则二级结构在折叠早期出现[16]。

二级结构是蛋白质折叠的基础，但对二级结构的预测仍然是一条“活恐龙”，经过30年的努力，准确率仍未达到80%。主要原因是序列上长程的残基对结构的影响难以估计。另一个也许更加基本的原因是，mRNA信息（包括密码子使用，mRNA局域结构等）可能对蛋白质二级结构的形成有一定影响。后者实际上是对Anfinsen由氨基酸序列决定蛋白质结构的原理的挑战。最近我们统计分析了现有序列结构资料，证明密码子的tRNA拷贝数和mRNA的stem/loop含量和三种蛋白质二级结构有十分明显的统计相关性。(Luo，Jia，Li，Biopolymers，2004[17]；Jia，Luo，Liu，Biopolymers，2004[18])。另外，我们还研

究了从二级结构序列决定蛋白质框架结构的问题(Luo，Li，Proteins，2000[19])。

从格子模拟中发现，序列决定构象必须经过中间环节－折叠途径，而折叠途径和动力学相关[20]。因此，在蛋白质折叠中和结构预测同样重要的是折叠途径的研究。

5. 生物功能网络问题

迅速增长的高通量实验资料给分子生物学提供了进行系统性分析的黄金机遇。一个系统并不等于其组成单元的简单加和。1+1≠2。生物系统分为蛋白质相互作用系统，核酸－蛋白质相互作用系统和脂－蛋白质相互作用系统三大类。任何生物功能皆须在一定的网络中展示。特征长度（平均最短距离，或称网络直径）和成团系数（结点的边数和最大可能值之比）是网络的两个静态几何量，规则网络有大的特征长度和成团系数，而随机网络正相反，特征长度和成团系数都小。在二者之间的是小世界网络，它同时具有小的特征长度和大的成团系数。小世界网络研究兴起后，生物学家开始投入到生物网络的研究中来。从代谢网络开始，到一般的蛋白质相互作用网络，到转录调节网络。他们发现和随机网络的 Poisson 分布不同，这些网络的度（连接）都遵从幂律分布，称为无标度性。仅很少结点和大量的边相连，而大多数的结点只有很少的连边。这是一种自组织现象。网络具有对攻击的耐受性，随机除去一些结点，网络的拓扑保持不变，而一旦具有最多连边的结点被删去，网络直径就迅速增加。例如酵母的蛋白质相互作用网络中，93%的蛋白连边数少于 5；0.7%的蛋白有 15 条以上的连边，它们在网络中起核心作用。这些网络一般都由若干模块按层次组成。模块性可以保证网络有较高的成团系数，并且成团系数与网络尺寸无关。层次模块组织有利于网络的鲁棒性。

分子生物学网络可以用拓扑模型或微分方程组的形式构建。由于资料的不完全和不确定，构建一个实际的网络并解出来，颇不容易。然而更深层次的问题是：建立在化学物理学基础上的网络可以描述生命的某些整体行为，但毕竟还不是生命。什么样的网络才是“活的”？这个问题和网络与环境的相互作用，和网络的热力学性质密切相关，是更具有挑战性的。

最后， 我愿用下面两句唐诗来结束这个讲话。

忽如一夜春风来，千树万树梨花开

江山代有才人出，李杜文章不新鲜

参 考 文 献

1. Kitano H. Systems Biology, “A brief overview,” *Science* 295, 1662-1664, 2002.
2. Sander C, http：//www.cbio.mskcc. org, *Memorial Sloan-Kettering Cancer Center,* 2004.
3. 罗辽复. 生命进化的物理观. *上海科技出版社,* 2000.
4. Gilbert W., “Towards a paradigm shift in biology,” *Nature* 349, 99-100, 1991.
5. Pullman B & Pullman A. Quantum Biochemistry. 1964.
6. Murray JD. Mathematical Biology. Berlin: *Springer-Verlag,* 1989.
7. Waterman MS. Introduction to Computational Biology. London: *Chapman & Hall,* 1995.
8. Schrodinger E. What is Life? (1944) (*Cambridge Univ Press*, 2001), 生命是什么？罗来鸥，罗辽复译，*湖南科技出版社,* 2003.
9. 罗辽复， “密码序列构象和动力学，” *大自然探索,* 7，9-14, 1988.
10. Luo LF, Li XQ., “Construction of genetic code from evolutionary stability,” *BioSystems* 65：81, 2002.

11. Luo LF, Li XQ., "Coding rules for amino acids in the genetic code - The genetic code is a minimal code of mutational deterioration," *Origins of Life* 32: 23, 2002.
12. Luo LF, Li H, Zhang LR., "ORF organization and gene recognition in the yeast genome," *Comp. Funct. Genome* 4: 318-328, 2003.
13. Zhang LR, Luo LF, "Splice site prediction with quadratic discriminant analysis using diversity measure," *Nucleic Acids Res.* 31: 6214-6220, 2003.
14. Luo, Lee WJ, Jia LJ, Ji FM & Tsai L., "Statistical correlation of nucleotides in a DNA sequence," *Physical Review E,* 58: 861-871, 1998.
15. Hsieh LC, Luo, Ji FM & Lee HC., "Minimal model for genome evolution and growth," *Phys. Rev. Lett.*, 90：018101-4, 2003.
16. Luo LF., "Conformation - transitional rate in protein folding," *Int J Quant Chem.*, 54：243, 1995.
17. Luo LF, Jia MW, Li XQ, "Protein structure preference, tRNA copy number and mRNA stem/loop content," *Biopolymers,* 73, 2004 May 20，published online.
18. Jia MW, Luo LF, Liu CQ., "Statistical correlation between protein secondary structure and messenger RNA stem-loop structure," *Biopolymers* 73: 16-26, 2004.
19. Luo LF, Li XQ., "Recognition and architecture of the framework structure of protein," *Proteins* 39:9-25, 2000.
20. Lee WJ, Luo LF., "A lattice model of nascent peptide folding," *Acta Sci. Natur. Univ. Intramongolicae,* 27:333 – 342, 1996.

作者简介：罗辽复，男，物理学教授，内蒙古大学博士生导师。早年从事粒子物理研究，1982 年后转向理论生物物理学，1987年后开始做生物信息学方面的工作。曾任中国生物物理学会理论生物物理学专业委员会主任。