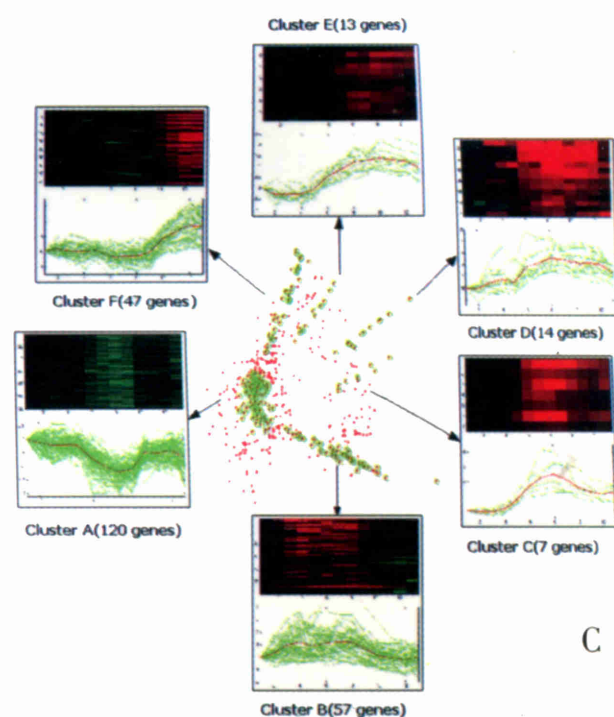
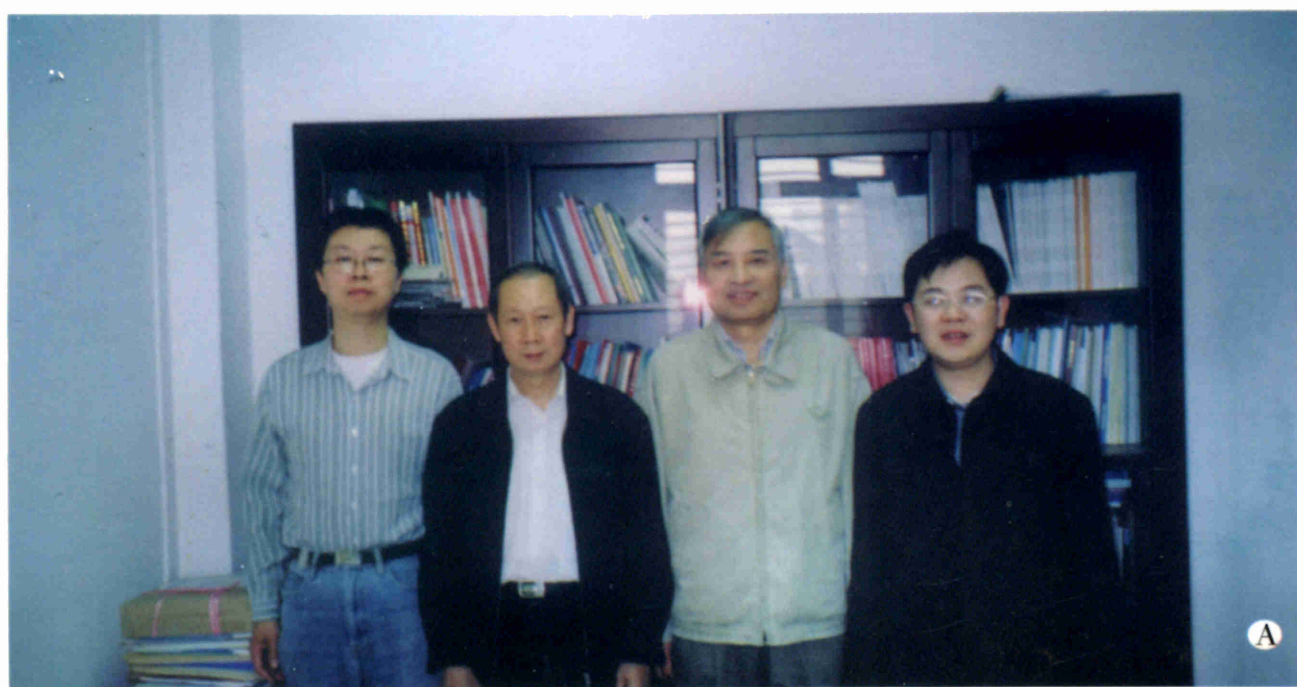


世界华人消化杂志®

**WORLD CHINESE
JOURNAL OF DIGESTOLOGY**

Shijie Huaren Xiaohua Zazhi

2003 年 10 月 15 日 第 11 卷 第 10 期 (Volume 11 Number 10)



10/2003

ISSN 1009-3079



名誉总编辑
潘伯荣
总编辑
马连生

World Journal of Gastroenterology® 被 SCI®-E, Research Alert®, Current Contents®/Clinical Medicine, Journal Citation Reports® Index Medicus, MEDLINE, Chemical Abstracts, EMBASE/Excerpta Medica 收录. 2002 年 JCR® 报告 WJG 影响因子 2.532. 世界华人消化杂志® 被 Chemical Abstracts, EMBASE/Excerpta Medica 收录. 2001 年中国科技期刊引证报告: 世界华人消化杂志® 影响因子 3.733, WJG 影响因子 2.920.

世界华人消化杂志

Shijie Huaren Xiaohua Zazhi

● 目 次 ●

2003 年 10 月 15 日 第 11 卷 第 10 期 (总第 114 期)

述 评	1465 复杂性疾病生物信息学研究的策略与方法 李梢, 张学工, 季梁, 李衍达
幽门螺杆菌	1470 幽门螺杆菌黏附素基因 babA ₂ 的克隆、序列测定及其生物信息学分析 白杨, 黄文, 王继德, 张兆山, 周殿元, 张亚历 1475 幽门螺杆菌 HspA 与大肠杆菌 LTB 基因融合及表达 郭红, 邹全明, 赵晓晏, 吴超 1480 人幽门螺杆菌热休克蛋白 A 编码基因的克隆、表达及抗原性研究 姜政, 蒲丹, 黄爱龙, 陶小红, 王丕龙 1485 幽门螺杆菌对克拉霉素耐药的分子基础 郝庆, 李岩, 高红, 张显忠
基础研究	1488 氧化苦参碱对四氯化碳诱导的大鼠肝纤维化 I, III, IV 型胶原表达的影响 陆伦根, 曾民德, 茅益民, 李继强, 邱德凯, 杨文卓, 贾一韬, 曹爱平 1492 粉防己碱、大黄与潘生丁抗肝纤维化作用比较 王如涛, 陈颖伟, 卫新革, 徐芹芳, 李定国 1497 珍珠梅水提物对大鼠肝损伤的保护作用 张学武, 朴龙, 刘超, 孙权, 金海玲, 尹宗柱 1500 乙型肝炎病毒 S 基因系列单突变克隆人工构建 余祖江, 杨东亮, 张俊, 郝友华, 王宝菊, 郝连杰 1505 急性胰腺炎大鼠肝脏 NF- κ B 对 ICAM-1 表达的调控及其意义 石力, 田伏洲, 黄大熔, 李旭, 赵碧, 顾大勇, 唐旭东, 王雨 1508 丁酸钠对结肠癌细胞株 HT-29 组织蛋白酶 D 表达水平的影响 李曦, 罗和生, 李凡 1511 国人青年结直肠癌解剖部位分布及临床病理特点 谢正勇, 卿三华 1515 慢性乙型肝炎病毒清除自杀基因平衡制约载体系统的构建 阙全程, 余祖江, 雷延昌, 杨东亮, 郝连杰 1520 人工构建含丙型肝炎病毒核糖体插入位点的双顺反子表达载体 阙全程, 余祖江, 雷延昌, 杨东亮, 郝连杰 1524 溃疡性结肠炎患者肠黏膜 Th1/Th2 类细胞因子 m-RNA 的表达 崔海宏, 陈村龙, 杨玉捷, 张祚建, 张耀东, 崔耀升
临床研究	1528 自膨胀金属支架治疗晚期食管癌吞咽困难 26 例 张朋彬, 赵晓晏, 李宜辉, 达四平 1531 胃癌组织 CD ₄₄ v9 和 MMP-2 基因的表达 张翠萍, 田宇彬, 赵清喜, 武军, 梁永信 1535 奥沙利铂综合治疗胃癌的疗效及机制 林万隆, 李定国, 陈强, 陆汉民, 马小明, 孙培龙 1540 聚合酶链反应检测 SEN 病毒 D 型和 H 型方法的建立及初步应用 唐蔚, 彭晓谋, 张瑛, 王辉, 蒋晓玲, 周伯平 1544 肝病患者血清 IGF-I 和 IGF-II 的变化 邵静鸣, 俞丽芬, 张曙, 吴云林 1547 ERCP 对儿童胰腺炎的诊断与治疗价值 李兆申, 许国铭, 施新岗, 邹晓平, 金震东, 孙振兴 1550 急性胆源性胰腺炎内镜诊治疗效及安全性 王东, 李兆申, 张文俊, 潘雪, 孙振兴, 邹晓平 1554 胰腺癌组织 ChAT, GAD65 和 PKC 酶活性的表达 杨竹林, 王群伟, 邓星辉, 李代强, 吕芳, 李永国 1558 国人胆囊结石的形态结构特征 吴杰, 杨海珉, 李静仪, 宋一德, 刘刚 1563 结核性腹膜炎与恶性腹水端粒酶活性 赵金满, 李福才, 于继红, 崔巍, 傅宝玉, 沙文阁
科研方法	1566 山莨菪碱联用地塞米松治疗腹部外科疾病并发 MODS 临床研究的操作方案 岳茂兴
文献综述	1569 门脉高压性肠病 尹朝晖, 刘浔阳 1572 肝纤维化治疗研究进展 叶方鹏, 肖冰, 张万岱 1576 现代肝脏局部解剖在活体部分肝移植应用的研究进展 方驰华, 朱新勇 1581 生长抑素类似物治疗肝细胞肝癌的抗肿瘤作用及其机制 冒海蕾, 黄介飞 1588 胰头部解剖在扩大胰十二指肠切除术中的应用 方驰华, 马俊勋, 钟世镇 1593 p53 基因在肿瘤基因治疗中的研究进展 张艳, 何凤田 1597 血管抑素的研究进展 陈建发, 黄宗海 1601 TGF β -Smad 信号转导通路与肝纤维化 吴晓玲, 曾维政, 王丕龙 1606 消化管发育中上皮细胞凋亡研究进展 李均, 汪维伟 1609 生物芯片技术及其在消化系统疾病研究中的应用 蒋业贵, 李兆申

文献综述	1614 Wilson病的诊断和治疗 林连捷, 郑长青 1618 E- 钙粘蛋白与食管癌侵袭转移的关系 吴静, 薛群基, 刘维民, 王爱勤, 寇伟 1621 胰腺癌的光动力学治疗 丁新民, 顾瑛, 刘凡光 1624 Ets 转录因子家族在发育和肿瘤发生中作用的研究进展 张健, 高福禄, 刘芝华 1628 核因子-κB 与细胞凋亡关系的研究进展 於亮亮, 于皆平, 罗和生, 于红刚
研究快报	1632 paxillin 在胃腺癌中的表达及临床意义 田素芳, 熊永炎, 余少平, 汪必成 1634 丹参对 TGF-β1 刺激的 NIH/3T3 细胞 <i>c-fos</i> mRNA 表达和 AP1 蛋白结合活性的影响 胡旭东, 王晓玲, 童普德, 吴小江, 刘平 1636 左旋精氨酸对大鼠肝脏缺血再灌注损伤的保护作用 郝悦, 周新民 1638 端粒酶在大肠癌细胞中的活性表达及临床意义 鲁明良, 林富林, 郑国宝, 姜朝晖 1640 多种因子在门脉高压大鼠结肠黏膜中的表达 尹朝晖, 刘浚阳, 黄飞舟, 黄穰浪, 任树平 1642 黄连素对 HT-29 人结肠癌细胞系 Ca ²⁺ 的抑制作用 台卫平, 罗和生 1645 DPC4 蛋白在不同病理分期的结肠肿瘤中的表达 唐朝晖, 邹声泉, 杨想平, 陈启奇 1646 Genistein 和 PD98059 对 aFGF 及 bFGF 诱导的 CCL229 细胞增生的抑制作用 尚海, 张颐, 单吉贤 1649 CO ₂ 气腹对肠道菌群生物学特性影响的实验研究 周丁华, 卫冰, 李宁, 黎介寿 1652 CO ₂ 气腹对大鼠胃肠肌电作用的实验研究 周丁华, 卫冰, 李宁, 黎介寿 1654 CO ₂ 气腹对胃黏膜血管活性肠肽及 P 物质含量的影响 周丁华, 卫冰, 李宁, 黎介寿
临床经验	1656 腹腔严重感染致多器官功能障碍的临床救治新对策 岳茂兴 1657 解毒固本冲剂治疗腹腔感染合并全身炎性反应综合征的临床研究 姜玉峰, 岳茂兴 1659 TIPSS 和 EVS 治疗食管静脉曲张破裂出血的临床分析 诸葛宇征, 王英德, 刘丽娜, 宫爱霞, 赵钢
消 息	1504 中国科技期刊走向世界的步伐正在加快 1568 欢迎订阅 2004 年度世界华人消化杂志 1571 欢迎订阅 2004 年度 World Journal of Gastroenterology® 1580 世界华人消化杂志获得 2001 年度百种中国杰出学术期刊 1613 WJG 搭建我国消化学基础和临床研究惟一国际交流的平台 1655 世界胃肠病学杂志英文版获得 2003-2004 年国家自然科学基金重点学术期刊专项基金资助
封面故事	1553 清华大学生物信息学研究所、生物信息学教育部重点实验室

世界华人消化杂志

Shijie Huaren Xiaohua Zazhi

吴阶平 题写封面刊名
陈可冀 题写版权刊名
(月刊)

创刊 1993-01-15
改刊 1998-01-25
出版 2003-10-15
原刊名 新消化病学杂志

总顾问 陈可冀
黄象谦
黄志强
黎介寿
刘耕陶
裘法祖
汤钊猷
王宝恩
危北海
吴孟超
吴咸中

社长总编辑 马连生
中文编辑 潘伯荣
王瑾晖
英文编辑 朱丽虹
排版 李少华
校对 李天华

张金哲
张学庸
赵东海
周殿元

编辑 世界华人消化杂志编辑委员会
030001, 山西省太原市双塔西街 77 号
E-mail: wcjd@wjgnet.com

出版 世界胃肠病学杂志社
100023, 北京市 2345 信箱
E-mail: wcjd @ wjgnet.com
http://www.wjgnet.com
电话: 010-85381892
传真: 010-85381893

印刷 北京科信印刷厂

发行 国内: 北京报刊发行局
国外: 中国国际图书贸易总公司
(100044, 北京 399 信箱)

订购 全国各地邮电局

邮购 世界胃肠病学杂志社发行部
(100023, 北京市 2345 信箱)
电话: 010-85381892
传真: 010-85381893
2003 年版权归世界胃肠病学杂志社所有

本刊已被国内外
检索系统收录

美国《化学文摘(CA)》
荷兰《医学文摘库/医学文摘(EM)》
俄罗斯《文摘杂志(PЖ)》
中国科技论文统计与分析
中国学术期刊文摘
中国中医药信息服务网
中国生物医学文献光盘数据库
《中文科技资料目录(医药卫生)》
中国生物医学期刊目次数据库
中国医学文摘外科学分册(英文版)
中国医学文摘内科学分册(英文版)

特别声明

本刊刊出的所有文章不代表世界胃肠病学杂志社和本刊编委会的观点, 除非特别声明. 本刊如有印装质量问题, 请向本刊编辑部调换.

ISSN 1009-3079	邮发代号	国外代号	国内定价	广告经营许可证
CN 14-1260/R	82-262	M 4481	每期 24.00 元 全年 288.00 元	1401004000050

www.wjgnet.com

复杂性疾病生物信息学研究的策略与方法

李 梢, 张学工, 季 梁, 李衍达

李梢, 张学工, 季梁, 李衍达, 清华大学生物信息学研究所、生物信息学教育部重点实验室 北京市 100084

李梢, 男, 1973-10-22 生, 安徽徽州人, 汉族. 1995 年北京中医药大学本科毕业, 2001 年北京中医药大学博士, 2001-2003 年清华大学生物信息学研究所博士后, 讲师. 主要从事自身免疫和炎症疾病、中医药生物信息学研究. 国家自然科学基金重点资助项目、青年基金项目, No.90209002, 30200365; 中国博士后科学基金资助项目, No. 2002-11

项目负责人: 李梢, 100084, 北京市海淀区清华园 1 号, 清华大学生物信息学研究所、生物信息学教育部重点实验室. shaoli@mail.tsinghua.edu.cn
电话: 010-62794294

收稿日期: 2003-06-06 接受日期: 2003-07-24

摘要

本文简述近年来复杂性疾病生物信息学研究的策略与方法, 并介绍清华大学生物信息学教育部重点实验室的有关工作. 由于遗传、环境的相互作用及基因型-表型复杂的内部结构, 常用的家系研究、基于遗传图谱的连锁分析、基于物理图谱的定位克隆以及关联分析等单基因病策略与方法, 在复杂性疾病的分子机制研究上存在局限. 在后基因组时代, 生物信息学的发展, 为从分子水平和系统观念研究复杂性疾病, 以及研究模式从“序列→结构→功能”向“相互作用→网络→功能”的转变提供了契机. 从多因素分析、基因的相互作用着手研究复杂性疾病成为热点. 我们以信息、系统的观点, 从功能基因组系统学出发研究复杂性疾病的机制, 并在复杂性疾病的基因组合及相互作用信息提取、复杂性疾病基因转录水平和表达水平的芯片分析、多层次生物信息整合、分子调控网络建模、中医药生物信息学等方面进行了有益的尝试.

李梢, 张学工, 季梁, 李衍达. 复杂性疾病生物信息学研究的策略与方法. 世界华人消化杂志 2003;11(10):1465-1469

<http://www.wjgnet.com/1009-3079/11/1465.asp>

0 引言

随着人类基因组计划的发展, 目前生命科学进入了快速、准确、低耗分析遗传和表达的信息时代. 生物信息学(Bioinformatics)作为当今生命科学研究最重要的平台技术, 其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^[1], 并阐明人类约 10 万种蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系, 寻求各种治疗和预防措施. 自 1980 年代末期以来, 生物信息学相继推动了系统生物学(systems biology)^[2]、功能基因组系统学^[1, 3]等的兴起, 发展了从系统观、信息结构、“复杂性”研究健康与疾病的新方法, 并已深入到与人类疾病密切相关的基因组学、蛋白质组学、药物基因组学^[4]等各个领域.

1 遗传-环境的相互作用及基因型-表型复杂的内部结构

疾病的发生与环境(环境有害因素)和遗传(遗传易感性)有关. 单基因病(monogenic disease)是疾病发病的遗传因素中以单基因缺陷占主导地位, 且在家系成员中疾病传递符合孟德尔规律者. 复杂性疾病(complex diseases)则是由多个基因及环境因素(包括致病微生物)相互作用所致, 且在家系中不符合孟德尔规律, 又称为多基因病(polygenic diseases), 多基因遗传病(polygenic inheritance diseases), 多因子病(multifactorial diseases). 如肿瘤、心血管病、代谢性疾病、神经-精神类疾病^[5]、免疫性疾病^[6]、帕金森症^[7]、肠癌^[8]等. 由于机体常见的疾病、健康状态是环境暴露、遗传易感性和年龄等因素复杂交互作用的结果, 基因-环境各种因素之间往往存在复杂的非线性关系, 因此导致研究上存在困难.

1.1 微效基因的困扰, 以及由此导致的预选基因难以选择 复杂性疾病的特点往往是由多个中效、微效基因共同决定疾病的复杂性状, 仅一个基因的改变对疾病的发生、发展, 以及对药物作用的影响不大, 其中某一或某些基因位点仅对应于该疾病的某个亚型、某个症状或体征^[9].

1.2 遗传与环境因素的交互作用分离 复杂性疾病的发病率, 即疾病外显性(penetrance)取决于后天环境因素影响的性质及程度. 目前按临床表现进行分类诊断的一个疾病, 实际上可能是由一组致病基因、或易感基因、或环境因素不同, 而表型相似的疾病组成, 如糖尿病. 群体中也存在具有遗传易感性但不发病或尚未发病的亚群. 缺乏对疾病形成过程中环境作用的有效控制途径, 必然导致研究对象的内部异质性, 从而影响疾病相关基因的研究. 随着多因素相关性疾病中遗传因素的作用以及遗传因素和环境因素互动认识的进一步深入, 将有可能改变现有的疾病分类方法.

1.3 疾病基因型与表型(gene-to-phenotype)存在复杂的关系, 以及丰富的内部结构^[10] 疾病基因型与表型存在多因素致病、多基因调控、涉及多个层次、临床表型复杂等特征. 如类风湿性关节炎(RA)的病理、免疫学改变虽然相似, 但其临床表现、基因表达谱改变却呈现出多样性^[11]. 同时, 复杂病的遗传易感性不一定是对疾病表型本身的直接影响, 而可仅是对疾病中间性状影响的间接后果. 可以认为, 现代医学与分子遗传学在疾病解释上存在可能的矛盾, 即对于复杂疾病而言, 分子遗传学的解释过于概括, 以致不适用于分为若干不同类别、乃至亚组的特定疾病.

2 几种利用遗传标记分析遗传病方法的局限

单基因遗传病由于因素单一, 基因型-表型关系清晰, 已经形成了较为成熟的患病家系研究、病例/对照分析法、患病同胞对法、传递不平衡测试法(transmission/disequilibrium test, TDT), 以及基于遗传图谱的连锁分析(linkage analysis)、基于群体病例-对照(case-control populations)的关联分析(association study)、基于物理图谱的定位克隆等策略与方法^[12, 13]. 其中遗传家系的连锁分析包括参数方法和非参数方法. TDT法实质上是一种以家系为基础的病例/对照关联分析法. NCB1的OMIM数据库已经收集了众多单基因遗传病的致病基因.

定位克隆已被证实是克隆基因尤其是人类群体中遗传疾病相关基因的一种有效手段^[14]. 复杂疾病的致病基因则很难用位置克隆法及致病基因定位、分离技术体系来确定^[15], 从可能的疾病易感位点到致病基因的跨越也充满困难. 相对于定位主效基因十分有效的连锁分析法来说, 关联分析法定位弱效基因的能力要强得多^[16]. 关联分析法侧重于研究遗传标记与疾病的关联关系, 然而受到需首先大致确定致病基因位置的局限. Baier et al^[17]曾试图利用位置克隆法定位II型糖尿病的易感基因. 运用关联分析表明, 一个位于内含子区域的多态位点UCSNP-43与疾病易感性密切相关, 这一位点位于CAPN10基因内. 对一段包含CAPN10基因的长为66 kb的区域进行测序发现了179个多态位点. 对其中的63个位点进行的病例/对照研究表明, 杂合体比纯合体更容易患病. 这意味着可能是CAPN10与其周围的某个遗传因素复杂的相互作用导致了糖尿病易感性的差异^[18]. 即是环境与人类基因、基因产物相互作用的结果, 以致最终的致病因素仍难确证.

3 新的研究策略: 生物信息学和系统观

从“序列→结构→功能”的生物学观点是既往复杂性疾病机制研究的基础, 即认为基因组本身包含蛋白质结构的所有必需信息, 这一观点过于简单并有太多还原论色彩. 由于生理-化学原理和生物学机制都不可避免地要涉及分子相互作用和反应的时空依赖性, 当前从“相互作用→网络→功能”的模式出发, 在转录组和蛋白质组层次研究基因调控网络、蛋白质相互作用网络的丰富信息成为后基因组学、生物信息学的前沿和热点^[19]. 同时, 功能基因组研究也开始朝复杂系统的方向发展^[1]. 这为复杂性疾病在大量已有数据资料的分析处理基础上, 由局部朝向整体, 由孤立朝向系统提供了可能.

3.1 以系统的观点适于阐发复杂疾病中遗传与环境的复杂关系 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. 基因组的调控与信息系统的调控具有相似的规律^[3]. 借鉴生命作为复杂系统的自组织现象, 以及基因的表达调控过程, 我们建立了基于基因表达调控的进化模型. 分析了其由简单结构向复杂结构进

化的趋势. 并发现高级复杂结构的出现是在一定的外界环境下, 生物种群内部出现的自组织过程^[20, 21].

3.2 以系统的观点适于阐发复杂疾病中基因功能的特点 计算与实验相结合的系统生物学方法有助于深入认识复杂的生物系统功能^[22]. 而在系统水平上理解生物体, 需要侧重于细胞、机体的功能而不是孤立的局部特点^[1]. 亦即研究各级产物的相互作用、整体形式比逐个研究基因产物将更为可取.

3.3 以系统的观点适于分析疾病基因型-表型之间的复杂关系 复杂疾病同一表型可能是由于同一代谢途径或信号传导途径上不同基因发生突变的结果, 在一个群体中孤立地研究某一个或几个热门基因将难以得出满意的结论^[23]. 因此, 在复杂疾病基因组分析中应注重对基因功能以及相关生化通路的阐明. 目前基因组研究已重视对生命基本过程的分析, 以及对基因调控、信号传导等的研究, 如生化途径数据库KEGG的建立^[24]等.

因此, 以系统观为特点的生物信息学研究策略与方法, 有望突破单基因病分析方法在复杂疾病研究中的局限. 并将在生物学和临床医学的诊断、治疗、药物开发等方面提供理论指导和分析, 具有广泛而关键的应用价值^[25-27].

4 复杂疾病的基因组合及相互作用信息提取

单核苷酸多态性(single nucleotide polymorphisms, SNPs)是疾病易感性、外显性、抵抗性以及药物反应性等生物学性状差别的重要遗传学基础. 很多疾病与基因突变或基因多态有关. 分析基因型数据, 特别是将SNPs数据与疾病和致病因素相关的计算方法是生物信息学研究的重点之一. 不同于传统的单基因研究方式, 目前对基因的相互作用、多个基因组合与疾病关系的研究日益受到关注.

在基因组层次上发展的多因子与表型关联算法, 可提供与表型相关的多个基因相互作用的信息. 复杂疾病的基因研究在设计上可表述为, 通过对病例和对照的实际样本(A, C, T, G)、一定数量SNPs的检测(可看作向量X), 从而寻找疾病有无或轻重的变量和向量X之间的关系. 现有的基于连锁分析、关联分析的研究多是针对X的单个分量展开, 忽略了各分量之间的关系. 提取有效基因组合的算法则是从各分量之间的关系入手, 研究多个可能致病因素的综合效果, 从而给疾病的预测、诊断以及个性化治疗提供依据. 我们应用小样本情况下的主效多因提取方法, 从基因的联合作用出发研究偏执型精神分裂症, 结果表明不同的SNPs组合可能反映出疾病的不同亚型^[28]. 同时, 在致病因素与临床表型的系上, 通过对胃癌前病变与Helicobacter pylori关系的Cluster分析表明, 不同临床表现的组合对于Helicobacter pylori感染具有不同的诊断准确率^[29]. 这为运用生物信息学对分子信息、病理及临床试验的资料进行整合, 通过一些重要的疾病致病因素, 从相互

关系阐释复杂疾病的病因及其诊断提供了初步的依据。

由于复杂疾病的机制涉及多个基因的累加作用, 以及某些环境因子的作用, 这些基因的SNPs及其特定组合可能是造成疾病易感性最重要的原因。最近Zhao et al^[30]提出了病例-对照研究中分析SNPs单倍型与环境因素的方法, 并通过分析载脂蛋白CIII(6SNPs)与冠心病术后再狭窄, 发现2种单倍型可能降低术后再狭窄患病风险。随着SNPs的不断发现和人类第三代遗传标记图的绘制等进展, 与疾病相关的人类基因信息数据库、与基因组多态性和基因突变有关的数据库dbSNP^[31]、HGBASE^[32]等相继建立, 为从SNPs探讨复杂性疾病的机制探讨提供了可行性。因此, 对疾病相关调节通路的候选基因进行SNPs的关联研究, 可能是多基因疾病研究取得突破的希望所在。目前我们同时还提出了一种新的基于SNPs数据, 用于研究复杂疾病和对疾病进行亚型分类的方法, 这一研究方法可以在小样本条件下充分利用已有的信息资源, 计算复杂度降低^[28]。

同时, 寻找与表型相关的有意义的多基因组合还需要加强以下方面的研究: (1)预选基因的采集需要集中于主要的代谢途径、调控通路, 从而避免由于样本分散导致的样本量相对减少; (2)实验设计中预选的疾病相关基因可能存在遗漏, 如果致病基因的作用具有累加性, 仍有可能准确地找到致病相关基因。候选基因组合中的最佳基因数目取决于样本数和基因缺失比率^[28]。因此研究中预先指定基因组合中的基因数目并不合理, 特别是孤立地研究单个基因与复杂疾病的关系往往不是最优选择, 并经常会导致信息利用的不充分。 (3)试验中病例本身的缺失也是潜在的偏差来源。目前对于缺失数据(missing data)的分析有均值、协方差的ML模型^[33], 自回归模型^[34], EM算法^[35], 归因(imputation)等方法, 然而仍需要通过进一步的研究, 形成较一致的认识。

5 芯片生物信息分析与多层次信息融合

解析复杂疾病基因型-表型的关系, 需要综合多层次信息, 如人类全部基因在染色体上的位置、序列特征(包括SNPs)、表达规律和产物(RNA和蛋白质)等。基因芯片, 如生物芯片(biochip)、微阵列(Microarray)、DNA芯片(DNA chip), 以及蛋白芯片等高通量检测技术, 使得高密度的数以万计的探针分子以及对杂交信号进行的检测分析变得切实可行, 是高效地大规模获取相关生物信息的主要手段。基因表达谱、蛋白质谱等芯片数据的生物信息学分析, 成为在基因转录、表达水平研究临床疾病的分子机制、疾病诊断和药物筛选的重要领域。目前模式识别、人工智能、统计学中的多种方法, 如聚类分析^[36], 人工神经网络^[37], 隐马尔可夫模型^[38]等, 已被用于在核酸和蛋白质两个层次上对基因表达谱, 即基因表达矩阵(行表示基因, 列表示实验样本)集合、分类及优化等信息的处理。并应用到肿瘤分型、肿瘤分类、基因功能研究、通路(pathway)分析、基

因调控网络构建、药物靶位识别等许多方面。对86例肺癌患者基因表达谱芯片的生物信息学分析发现数百个基因与癌细胞显著相关, 属于细胞分裂、蛋白降解、嘧啶与嘌呤代谢、氧化磷酸化等通路^[39], 从信号通路调节异常的分析中则可预测肿瘤表型^[40]。在神经网络等传统方法得到很多应用的同时, 统计学习理论(statistical learning theory, SLT), 即专门用于研究小样本情况下机器学习规律的理论日益受到重视, 以支持向量机(SVM)^[41]方法为代表的SLT方法将为芯片的生物信息学分析提供新的途径, 从而进一步推动人们在基因组水平上以系统的、全局的观念去研究疾病多层次的现象及其本质。

6 分子调控网络建模

基因的结构或种类决定物种, 基因的功能或表达则决定生命的健康、疾病等状态。基因调控网络、蛋白调控网络、代谢网络等分子调控网络的研究, 是理解基因组功能、进而理解复杂疾病本质正在发展的方向, 并将为复杂性疾病多基因之间的关系、基因型-表型各部分、各层次的相互作用、调控通路提供背景和依据^[42]。其中在转录组层次上发展的基因调控网络建模方法可提供与表型相关基因调控网络的部分信息。在Promoter区的调控元件可提供基因之间相互作用的线索。随着基因组数据的指数式增长, 有关生理、病理数据库快速发展, 如基因表达有关的数据库^[43], 蛋白质序列数据库有PIR^[44], Swiss-Prot^[45]; 蛋白质结构数据库有PDB^[46]等, 以及选择性剪接数据库^[47]等。数据的大量积累也使得从系统水平上研究分子调控网络成为可能。

对于基因调控、细胞信号传导过程等已具有大量的分子生物学实验研究, 急需对实验结果进行理论总结。基因、蛋白等调控网络的建模, 即综合生理病理相关的基因、SNPs、基因表达状况、蛋白质功能状态以及临床治疗等信息, 以系统方法将不同的测量数据、各种因素的辨识, 各个层次上的相互作用关系进行整合, 利用数学模型深入了解生物的结构和功能以及生理病理过程。目前分子调控网络目前大多为简单系统、局部网络的基础研究, 以及从大规模基因表达谱中发现和描述基因调控网络的研究。基因调控网络建模主要在于代谢网络^[48]、信号通路网络^[49]、生化网络^[50]、运用数据挖掘方法构建人类基因表达网络^[51]、运用多元回归分析构建DNA微阵列数据基础上构建基因调控网络^[52]、神经网络^[53]等。按控制论的基本原理, 我们曾归纳出基因调控过程中的7种典型控制环节, 基因调控和信号传导路径可以由这些控制环节组合而成, 控制论的研究有助于寻找其中的共同规律^[54]。通过基因组、转录组等层次多种信息相互印证, 则有望扩充已知基因调控网络的方法。

随着系统建模各种手段的发展, 尤其是复杂系统仿真、建模等方法的建立, 表明目前已有能力与方法

将复杂的生物系统问题“降维”处理。依据一个复杂系统的输入(遗传物质的差异)和输出(基因表达谱、蛋白质谱)确定各种变量之间的因果关系,有望逐步建立与疾病密切相关的分子调控网络,从而最终解释复杂疾病的成因、治疗等各种理论问题。也将为降低复杂疾病有关信息分析的难度,减少实际观测样本量的限制提供依据。

7 中医药生物信息学

中医药学植根于临床实践,防治众多常见疾病具有较佳的疗效,是世界医学不可或缺的组成部分。然而长期以来研究策略与方法的局限,制约了中医药治疗规律和疗效机制的探讨。将侧重于宏观的中医药学及其诊疗特色、丰富资源与侧重于微观的生物信息学进行有机结合,开展“中医药生物信息学”(traditional chinese medicine bioinformatics, TCMB)有关理论与方法的研究,是我们提出的一个充满挑战和机遇的方向^[55]。我们的研究表明,以生物信息学的有关理论与方法为桥梁,在中医学研究中发展相应的信息分析与整合手段,将有助于深入了解以整体观、辨证论治为核心的中医诊疗规律;一套行之有效的中医药生物信息分析方法,也将为在分子水平上发掘中医药的系统内涵,在系统层次上加深对于复杂性疾病的理解提供新的途径。

7.1 中医证候生物信息分析 由于中医学对人体体质因素、环境因素及疾病不同阶段证候演变的整体认识方法,使得在微观层面上研究中医学诊疗机制,必然需要大量的生物信息分析,特别是非线性规律的分析尤为突出。中医学对疾病的观察体现出机体与环境相交融的整体观念。对空间的“证”、时间的“候”进行证候学判断,从而指导方药的治疗则是中医临床的核心特点。我们运用控制论方法与中医理论相结合,已初步揭示在机体与环境、机体内多因素相互作用基础上的稳态机制和系统特性^[56, 57]。通过“降维升阶”等处理,则有助于从复杂的四诊信息、理化信息以及多层次生物信息中,使主要证候因素得以辨识^[58, 59]。目前国际上认为,揭示复杂疾病机制的新途径,在于了解亚细胞、细胞、组织、器官及系统结构中的蛋白质相互作用,以及认识基于相互作用的疾病不同状态^[60],而多学科的协作、有效的信息整合,其意义超过了单纯的基因研究^[61]。我们开展的中医证候与复杂性疾病有关病理、生理多层次信息的综合研究^[62],也符合了这一发展趋势。

7.2 中药药效系统学研究 中药成分相对复杂,治疗疾病时往往通过不同环节发挥整体调节作用^[59]。针对方剂、证候高维小样本数据,我们建立了统计模式识别分析的PARM算法。应用于评价中药不同配比的分析与优化、寒热方剂的作用、寒热证候的特征等,取得了较好效果。与统计学方法相比,该模型可合理利用启发式的直观信息,适合于药效的特征提取、系统评价,而不仅是给出差异性的比较结果。表明“中药药效系统学”

是符合中药作用特色的研究途径。

7.3 中医药与复杂疾病研究的互补 中医药在治疗上注重功能调节,可能是在调控疾病的相关(易感)基因的表达及表达产物上发挥重要作用,已发现中药可以影响一些细胞因子、组织损伤酶的比例^[63]、动态变化等。中药的作用具有多因微效基础上的突现特点,在一定意义上符合多基因病的形成特征及治疗趋势。因此,充分发挥生物信息学计算、设计的作用,在计算与实验相结合的研究框架下,进行基于多层次生物信息分析的中药有效分子组合筛选,中药生物信息与化学信息相关性研究,以及中药基因组学等研究,对于复杂疾病的药物设计及防治具有较大价值;同时,以确有疗效、成分明确的中药为探针,可望开启复杂疾病机制研究中“以药测病”的思路,从新的角度扩充疾病相关的分子调控网络^[62],并促进生物信息、医学信息的融合。

总之,生物信息学掌握着基因组的真正用途,并将引起生物医学以及临床诊疗的革命性变化^[64]。作为当今生命科学研究最重要的平台技术,生物信息学不仅能够分析复杂疾病多种生物分子数据,同时更适于综合多种生物分子及其相互作用的知识来了解生物系统的功能,由获取有关生物知识迈进到对生命基本规律的认识,并促进复杂性疾病的研究向功能、系统的方向发展。

8 参考文献

- 1 李衍达. 以信息系统的观点了解基因组. 电子学报 2001;29:1731-1734
- 2 Kitano H. Systems biology: a brief overview. *Science* 2002; 295:1662-1664
- 3 李衍达. 信息与生命. 化学通报 2001;10:601-607
- 4 Bayat A. Science, medicine, and the future: Bioinformatics. *BMJ* 2002;324:1018-1022
- 5 Bray NJ, Owen MJ. Searching for schizophrenia genes. *Trends Mol Med* 2001;7:169-174
- 6 Ulfgren AK, Grondal L, Lindblad S, Khademi M, Johnell O, Klareskog L, Andersson U. Interindividual and intra-articular variation of proinflammatory cytokines in patients with rheumatoid arthritis: potential implications for treatment. *Ann Rheum Dis* 2000;59:439-447
- 7 Warner TT, Schapira AH. Genetic and environmental factors in the cause of Parkinson's disease. *Ann Neurol* 2003;53(Suppl 3):S16- S23
- 8 Augenlicht LH, Heerdt BG, Mariadason JM, Yang W, Wilson AJ, Fragale A, Velcich A. Environment-gene interactions in intestinal cancer. *Eur J Cancer Prev* 2002;11(Suppl 2):S12-S17
- 9 Felsenfeld S. Finding susceptibility genes for developmental disorders of speech: the long and winding road. *J Commun Disord* 2002;35:329-345
- 10 Lazzaroni LC, Karlovich CA. Genotype to phenotype: associations, errors and complexity. *Trends Genet* 2002;18: 283-284
- 11 Kamradt T, Mitchison NA. Tolerance and autoimmunity. *N Engl J Med* 2001;344:655-664
- 12 Sun FZ, Flanders WD, Yang QH, Zhao HY. Transmission/disequilibrium tests for quantitative traits. *Ann Hum Genet* 2000;64(Pt 6):555-565
- 13 Zhao H. Family-based association studies. *Stat Methods Med Res* 2000;9:563-587
- 14 Collins FS. Positional cloning: let's not call it reverse anymore. *Nat Genet* 1992;1:3-6

- 15 Patterson M. That damned elusive polygene. *Nat Rev Genet* 2000;1:86
- 16 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516-1517
- 17 Baier LJ, Permana PA, Yang X, Pratley RE, Hanson RL, Shen GQ, Mott D, Knowler WC, Cox NJ, Horikawa Y, Oda N, Bell GI, Bogardus C. A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. *J Clin Invest* 2000;106:R69-R73
- 18 Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000;26:163-175
- 19 Kanehisa M 著. 孙之荣译. Post-genome Informatics. 第1版. 北京: 清华大学出版社、牛津大学出版社, 2002:114
- 20 周晴, 计宏凯, 李衍达. 基于基因表达调控的进化模型. 清华大学学报(自然科学版) 2002;42:135-139
- 21 卢欣, 孙之荣, 李衍达. 基因组复杂度进化的仿真研究. 生物物理学报 2001;17:318-328
- 22 Kitano H. Computational systems biology. *Nature* 2002;420:206-210
- 23 Emahazion T, Jobs M, Howell WM, Siegfried M, Wyoni PI, Prince JA, Brookes AJ. Identification of 167 polymorphisms in 88 genes from candidate neurodegeneration pathways. *Gene* 1999;238:315-324
- 24 Kanehisa M. The KEGG database. *Novartis Found Symp* 2002;247:91-101
- 25 Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* 2001;291:1224-1229
- 26 Butler D. Genomics. Are you ready for the revolution? *Nature* 2001;409:758-760
- 27 Debouck C, Metcalf B. The impact of genomics on drug discovery. *Annu Rev Pharmacol Toxicol* 2000;40:193-207
- 28 计宏凯, 季梁. 基于 SNPs 的复杂疾病基因分析和亚型分类. 第二届中国生物信息学大会论文集. 北京: 2002:6
- 29 Li S, Lu AP, Zhang L, Li YD. Anti-*Helicobacter pylori* immunoglobulin G (IgG) and IgA antibody responses and the value of clinical presentations in diagnosis of *H. pylori* infection in patients with precancerous lesions. *World J Gastroenterol* 2003;9:755-758
- 30 Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet* 2003;72:1231-1250
- 31 Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000;28:352-355
- 32 Brookes AJ, Lehtvaslaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, Bork P, Ortigao F. HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res* 2000;28:356-360
- 33 Horton NJ, Laird NM. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics* 2001;57:34-42
- 34 Albert PS, Follmann DA, Wang SA, Suh EB. A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics* 2002;58:631-642
- 35 Albert PS. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* 2000;56:602-608
- 36 Lee JM, Sonnhammer EL. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* 2003;13:875-882
- 37 O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics* 2003;4:13
- 38 Ji X, Li LJ, Sun Z. Mining gene expression data using a novel approach based on hidden Markov models. *FEBS Lett* 2003;542:125-131
- 39 Creighton C, Hanash S, Beer D. Gene expression patterns define pathways correlated with loss of differentiation in lung adenocarcinomas. *FEBS Lett* 2003;540:167-170
- 40 Graeber TG, Eisenberg D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat Genet* 2001;29:295-300
- 41 张学工. 关于统计学习理论与支持向量机. 自动化学报 2000;26:32-43
- 42 Huang S. Rational drug discovery: what can we learn from regulatory networks? *Drug Discov Today* 2002;7(20 Suppl):S163-S169
- 43 Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000;24:236-244
- 44 Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao C, Yeh LS, Ledley RS, Janda JF, Pfeiffer F, Mewes HW, Tsugita A, Wu C. The protein information resource (PIR). *Nucleic Acids Res* 2000;28:41-44
- 45 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilboud S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365-370
- 46 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235-242
- 47 Ji H, Zhou Q, Wen F, Xia H, Lu X, Li Y. AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res* 2001;29:260-263
- 48 Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000;406:651-655
- 49 Karp PD. Pathway databases: A case study in computational symbolic theories. *Science* 2001;293:2040-2044
- 50 Kitami T, Nadeau JH. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nat Genet* 2002;32:191-194
- 51 Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;28:21-29
- 52 Kato M, Tsunoda T, Takagi T. Inferring genetic networks from DNA microarray data by multiple regression analysis. *Genome Inform Ser Workshop Genome Inform* 2000;11:118-128
- 53 Pouget A, Latham P. Digitized neural networks: long-term stability from forgetful neurons. *Nat Neurosci* 2002;5:709-710
- 54 卢欣, 李衍达. 基因调控过程中的典型控制环节. 自动化学报 2000;26:638-645
- 55 Li S. Advances in TCM symptomatology of rheumatoid arthritis. *J Tradit Chin Med* 2002;22:137-142
- 56 庄永龙, 李梢, 李衍达. 基于控制论的中医学四时五脏系统稳态性能仿真. 系统仿真学报 2003;15:922-926
- 57 李梢, 王永炎, 季梁, 李衍达. 复杂系统意义下的中医药学及其案例研究. 系统仿真学报 2002;14:1429-1403
- 58 李梢. 从维度与阶度探讨中医证候的特征及标准化方法. 北京中医药大学学报 2003;26:1-4
- 59 Li S, Lu AP, Jia HW. Therapeutic actions of the Chinese herbal formulae with cold and heat properties and their effects on ultrastructures of synoviocytes in rats of the collagen-induced arthritis. *J Tradit Chin Med* 2002;22:290-296
- 60 Noble D. Modeling the heart-from genes to cells to the whole organ. *Science* 2002;295:1678-1682
- 61 Kiberstis P, Roberts L. It's not just the genes. *Science* 2002;296:685
- 62 李梢. 中医证候与分子网络调节机制的可能关联. 中国科学技术协会首届学术年会报告. 周光召. 面向21世纪的科技进展与社会经济发展, 第1版. 北京: 中国科学技术出版社, 1999:442
- 63 Li S, Lu AP, Wang YY, Li YD. Suppressive effects of a Chinese herbal medicine Qing-Luo-Yin extract on the angiogenesis of collagen induced arthritis in rats. *Am J Chin Med* 2003;31:861-866
- 64 Kim JH. Bioinformatics and genomic medicine. *Genet Med* 2002;4(6 Suppl):62S-65S



Published by **Baishideng Publishing Group Inc**
7901 Stoneridge Drive, Suite 501, Pleasanton,
CA 94588, USA
Fax: +1-925-223-8242
Telephone: +1-925-223-8243
E-mail: bpgoffice@wjgnet.com
<http://www.wjgnet.com>



ISSN 1009-3079

