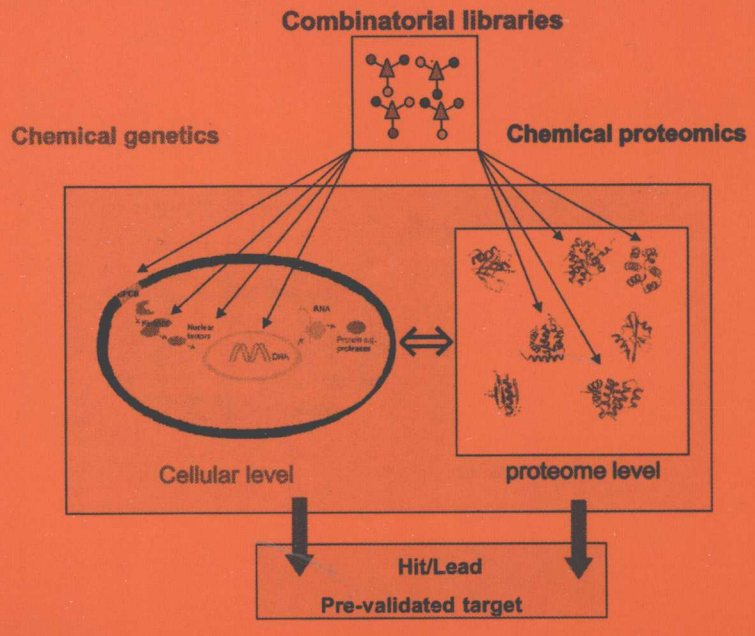
现代化学前沿译丛

化学基因组学

〔美〕Ferenc Darvas
〔美〕András Guttman 主编
〔匈〕György Dormán

杨振军 等 译



科学出版社
www.sciencep.com

现代化学前沿译丛

化学基因组学

〔美〕Ferenc Darvas
〔美〕András Guttman
〔匈〕György Dormán
主编

杨振军 等
译

科学出版社
北京

图字：01-2006-6133

内 容 简 介

本书由该领域的国际知名专家撰写，是第一本全面的化学基因组学专著，全面讨论了化学基因组学的各个专题，以及从相关计算机芯片到实验等的相关技术。第一部分描述了化学基因组学的定义和基本概念，包括从计算机芯片化学基因组学到基于芯片/微阵列等的主要技术。第二部分专注于特殊技术，讨论了小分子探针研究特异基因产物方法的出现和应用实例。最后3章是药物发现相关领域的研究实例，进一步强调了化学基因组学是多个研究方向相互关联的综合性研究方法，如脂质组学、神经递质的药学基因组学方向，以及DNA专利的近似线性配对算法。

本书可供药物化学、分子生物学、生物信息学等专业科研人员参考，也可作为药物化学专业教材使用。

图书在版编目(CIP)数据

化学基因组学（Chemical Genomics）/（美）达维斯（Darvas, F.）等主编；杨振军等译. —北京：科学出版社，2008

（现代化学前沿译丛）

ISBN 978-7-03-020560-5

Ⅰ. 化… Ⅱ. ①达…②杨… Ⅲ. 化学-基因组 Ⅳ. Q343.1

中国版本图书馆CIP数据核字（2008）第031099号

责任编辑：杨 震 黄 海/责任校对：包志虹

责任印制：钱玉芬/封面设计：陈 敬

科 学 出 版 社出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

铭浩彩色印装有限公司印刷

科学出版社发行 各地新华书店经销

*

2008年4月第 一 版 开本：B5（720×1000）

2008年4月第一次印刷 印张：13 1/2

印数：1—3 000 字数：260 000

定价：50.00元

（如有印装质量问题，我社负责调换〈明辉〉）

前　言

在后基因组时代，有希望出现3000～10 000个新的疾病相关的药物作用靶。药物研究的一个重大挑战是将这个领域如此多的研究成果转化为新药。有前景的是，这些新的药物作用靶既可以针对传统疾病，又可以为新出现的疾病提供新的治疗方案。

在人类基因组计划完成了测序阶段并绘制了人类生命的蓝图后，现在研究工作的重点是揭示基因功能，如利用功能基因编码的蛋白质加以验证。随着天然组合化学的开展，利用化合物库的手段可以扩展到利用小分子干预整个基因组或蛋白质组。正如Schreiber指出的，这个领域的一个前沿性研究工作就是针对每一个基因产物，利用小分子伴侣研究有机体和细胞的功能，这也就是化学基因组学的研究目标。化学基因组学确实担负了完成这一探寻的重任，例如，利用小分子与某一个特定基因表达的一个或多个蛋白质的相互作用，确定这一基因的功能和生物学作用。此外，各种化学基因组学方法可以阐述那些遗传操作方法或结构基因组手段不能解决的问题。利用小分子在细胞水平的筛选来发现生物活性分子的理念，是药物化学和药理学的中心工作。通过整合特殊的基于亲和性评价技术和高通量筛选的方法，化学基因组学引导的技术能够应用新的平行研发手段，在一个单步实验中找到药物作用靶和中标化合物，靶的确证和先导化合物的优化可以同时完成，这极大地促进了临床前药物的开发速度。在过去的两三年里，基因组学带动了药物发现、化学信息学和高通量筛选等领域的非常有意义的进展，文献中利用化学基因组学工具完成的工作所发表文章数量的增加有力地证明了这一点。

本书由该领域的国际知名专家撰写，是第一本全面的化学基因组学专著，全面讨论了化学基因组学的各个专题，以及从相关计算机芯片到实验等的相关技术。第一部分描述了化学基因组学的定义和基本概念，包括从计算机芯片化学基因组学到基于芯片/微阵列等的主要技术。第二部分专注于特殊技术，讨论了小分子探针研究特异基因产物方法的出现和应用实例。最后3章是药物发现相关领域的研究实例，进一步强调了化学基因组学是多个研究方向相互关联的综合性研究方法，如脂质组学、神经递质的药学基因组学方向，以及DNA专利的近似线性配对算法。

总之，这本书在这个新出现的领域，前瞻性地为药物发现、药物化学、分子生物学以及化学和生物信息学等领域的专家们做了综合性的概括。本书对于在实

验室工作的科学家及研究生、本科生也是非常有用的，同时，也为大学教授在后基因组时代开设一门药物化学专业课程提供了一本有价值的教材。

Ferenc Darvas
András Guttman
György Dormán
杨振军 译

目　录

第 1 章　什么是化学基因组学？

PAUL R. CARON
Vertex Pharmaceuticals Inc., Cambridge, Massachusetts, U. S. A.
乔任平 译，杨振军 校译

对于生物学的探索不仅要靠科学探索精神，更需要生物或者化学工具的操作系统的发展来推动。在 20 世纪后半叶，生物系统的研究已经从细胞水平上升到分子水平。很多生物（包括人类）的全基因组测序的完成，使人们对生物的理解产生了翻天覆地的变化。现在，着眼于生物体并在器官、组织、细胞、亚细胞及分子水平进行剖析已经不是人们关注的焦点，因为虽然已经在分子水平的研究上达到了相当精密的程度，但是关于这些分子之间的相互作用，以及如何整合在一起作为功能单位的机制还知之甚少。

很多基因水平上的数据和生物学工具在应用上是很有价值的，包括小分子干扰 RNA（siRNA）、反义技术（antisense）、基因敲除技术（knock-out）、转基因技术（transgenics）、抗体技术（antibody）及群体遗传学（population genetics）。这些技术手段可以针对一个靶基因，也可以针对多个靶基因应用。然而，人类基因组中有约 30 000 个基因，上述技术不能很好地阐明每个基因的生物功能，同时耗时长，成本昂贵，以及耗费时间和资源。因此，这些生物学工具更基本的应用可能是在基因组子集水平的研究上，以提供充分的数据进行计算并验证相关理论的可靠性。目前对于基因组的许多基因，通过干扰相关基因只能进行很有限地预测分析并推断某个基因的功能。

在应用这些技术确证已知药物的特定蛋白靶标前，药学和基因组学长久以来都是两个相对独立的领域。对于药物与靶点之间相互作用的研究，以及在分子水平上逐步深入地理解靶点在特定生物途径中的作用角色，已使目前市场上的药品数量和性质都有很大的突破。然而，目前基因组中的基因只有很少一部分作为药物的靶标。虽然干扰基因组中一些基因之间的相互作用并不一定会带来治疗上的生理后果，但是并不代表全部重要的药物靶点都已经被确证了。

一些新的药物靶点可以通过一些生物分子工具和靶向特异性化合物来确定。虽然药物靶点的基因调控很少能够模拟出化学激动剂或拮抗剂对靶基因引起的精

确影响，但是仍然可以为相关的生物学途径的研究提供线索。有很多方法可以用来确定化合物，这些化合物需具有适当水平的、对于靶标起作用的活性和专一性，以指定靶标的生物学功能，例如，基于天然产物衍生物的高通量筛选、组合化学，基于同一个母核的相关衍生物的合成化学。此外，结合了基于结构的药物设计和分子模拟的传统药物化学及化学，也可以用来进一步优化这些化合物。

在当前基因组时代，具有挑战性的已经不是研究单个靶标或者优先研究药物干预可能产生治疗作用，而是尽快地赋予已经破译序列的基因以生物学功能。在过去的十年中，技术的突破和信息学领域的进步可以直接应用到这些研究工作中。

一个新的领域总是会有一些新的名词出现，并且会有一些关于它们的准确定义的争论。在这本书里我们是这样定义的：化学遗传学（Chemical genetics）是利用作用活性高的、有选择性的化合物来研究生物途径的过程；化学基因组学（Chemical genomics）是研究处理广大多重靶标的过程，这比单独研究一系列的不相关的靶标更有效率；化学基因组学（Chemogenomics）是结合化学基因组学、化学遗传学及信息学工具，针对多重靶标鉴别类药分子。

通过化学生物学方法有许多途径可以实现更高的效率。着眼于一个特定的基因家族可以再次利用基因克隆、表达、纯化和将一个给定的靶标特性化的专业技术促进多重靶标的研究。很多基因家族的成员拥有相同的用于体外分析的底物或者天然配体。例如，离子通道和G蛋白偶联受体拥有共同的细胞信号转导通路，这就大大减小了对这些分子家族确立进一步分析目标的必要性。事实上，对于任何一个靶标都会建立多重计数筛选体系。化学生物学是在此基础上的延伸，它不仅仅像计数筛选体系作为一个已知化合物的“被动滤器”，更重要的是一个在这些可替代的评价体系中鉴定化合物对药物靶标起拮抗作用的主动过程。

化学生物学另外一个很大的优势在于可以将化合物作为基于同一个母核的相关衍生物的组合化学的一部分，或者作为针对和其他基因相关的某个基因的潜在先导化合物分子。当这些给定基因家族的分子文库被建立起来以后，接下来药物先导物的数量和性质能力都将得到大大的提高。另外，由于这些文库大多由化学可控的分子组成，对于一些相似的化学反应可以进行类似的进一步优化。这些领域的机构经常将大量的精力投放在这些分子的性质研究上，并为此制定了清晰的研发策略。

选择出来的分子可以用来确定给定基因（化学遗传学）的生物角色及其治疗潜力。显然，这种发现分子的过程越有效率，就可以评估越多的靶点。以前，要花费很多时间和精力通过动物模型寻找具有适宜的药动学和药效学参数的分子。适当的应用化学生物学手段可以减少这方面研究的困难，这些生物学方面的探索可以使研发过程变得比以前更容易进行。

Chemical genomics（化学基因组学）与 Chemogenomics（化学基因组学）的区别是，前者的影响在于提高实验工作的效率，而后者包括一些信息学成分，将实验中所得结论利用外推法应用到其他的基因家族上。化学基因组学需要将化合物结合到靶标及其同源类似物的三维结构上和（或）建立与相关基因家族相对应的化合物库。精细的同源模建及构效分析可以用来分析针对某个新的基因靶标的、具有一定潜力的选择性抑制剂。

第 2 章　芯片上的化学基因组学

CYÖRGY M. KESERŰ and ZOLTÁN KOVÁRI
Gedeon Richter Ltd., Budapest, Hungary
乔任平 译，杨振军 校译

2.1　引言

化学基因组学利用小分子作为探针来探究生物学途径。本文将探讨小分子与基因产物是否相互作用的问题。蛋白质与配体的相互作用可以广泛地运用计算机手段（如使用芯片技术）进行研究。在化学基因组学的计算机研究中，研究者需要原子水平上可分辨的感兴趣蛋白质的 3D 结构。蛋白质结构通过实验或者理论计算得到，用来进行原子水平上的蛋白质-配体相互作用的研究。本章我们会对将计算机技术用于化学基因组学研究的可能性进行概括论述。该项技术手段的瓶颈是蛋白质模型的质量能否达到要求，因此我们首先来对如何获得蛋白质模型进行一个比较详细的描述。因为在各类文献中将计算机技术应用到蛋白质-配体相互作用已有很详细的论述，这里我们就不再阐述这类相互作用的具体细则。取而代之，我们将就化学基因组学相关的蛋白质数据库（如序列及结构数据库）怎样运用到蛋白质模建技术及蛋白质功能预测的方法学进行深入的讨论。

2.2　将计算机技术应用到化学基因组学

2.2.1　蛋白质结构

原子水平上蛋白质-配体相互作用的计算机模拟研究，需要知道蛋白质原子水平可分辨的 3D 结构。这样的 3D 结构通过实验或者理论计算获得。目前，蛋白质同等物的 3D 结构主要以 Protein Data Bank（PDB 文件格式）的格式给出，这种格式可以被目前大多数软件解读。

1. 实验的蛋白质结构

通过实验方法可以得到优势的蛋白质 3D 结构，如 X 射线晶体衍射（X-ray crystallography）或者核磁共振技术（NMR），两者各有优势和局限性。

虽然蛋白质晶体衍射手段不受分子质量的限制，但是蛋白质的结晶能力却是绝对必要的。蛋白质的可结晶性是该技术的瓶颈。要完成结晶必须尝试几个条件的变化，即蛋白质浓度、沉淀剂、pH 及温度。晶体化对于存在于两相中的蛋白质来说是很困难的，如跨膜蛋白既有亲水区（处于膜外）又有疏水区（处于膜内）。可以想像，要想把这类蛋白质分子有序排列成为晶体化不是一件容易的事情。G 蛋白偶联受体家族（GPCR）拥有 7 个跨膜螺旋结构，这对于蛋白质结晶来说是一个挑战，它们当中只有一个（牛视紫红质蛋白，bovine rhodopsin[1]）被成功结晶。然而，大约有 50%被批准的药物通过该蛋白家族发挥作用[2]，使得该家族的原子结构具有研究意义。如果配体结合区不是在跨膜区域，那么一种用来晶体化的方法就是切断并结晶可溶区域。这种方法已经成功地应用在 APP 切割酶的 β 位点（BACE）[3]及 GluR2 受体[4]。当晶体形成，蛋白质进入一种固体状态，这有别于蛋白质发挥正常功能的自然状态。这种晶体蛋白有非自然的、相当高的浓度，并且晶体之间的联系防止了大规模的区域运动。虽然具有上述的局限性，X 射线衍射还是对蛋白质和蛋白质-配体复合体的原子解析结构给出了大量的信息。因为晶体结构主要呈现蛋白质的溶液态折叠以及配体结合的几何学特征，该方法的程序比较简易，使得研究者希望阐明蛋白质结构时首选 X 射线衍射法。

当用这些晶体结构进行进一步的计算机水平上的研究时，不仅仅需要配合物，还需要某些结构区域的特征数据。关于晶体结构的一个很重要的参数就是分辨率，高的分辨率意味着低的数值（如 1.8Å 分辨率的晶体结构要比 2.3Å 的准确）。对于计算机计算而言，推荐使用可能得到的最高分辨率，实际上高于 2.5Å 的分辨率就是可以接受的了。

可靠性因子（R-factor）是晶体结构的很重要的指标。虽然晶体学中使用了几种 R 因子，其中一个“R”（没有任何指数）是最重要的一项。该因子描述了所得结构与实验得到的衍射形态的吻合程度。R 因子越低，所得结构的可靠性就越好，通常 R 因子低于 20%的结构可以作为研究之用。

晶体结构另外一个重要的特征是原子和残基的可动性，用 B 因子（B-factor）来描述。在高解析结构中，B 因子是针对单个原子而言的；而在低解析结构中，一项平均 B 因子是对原子基团而言的。B 因子是 PDB 坐标文件的一部分，以 $Å^2$ 为单位表征原子或基团的可动性，也就是预测原子出现在它的坐标的可能性。B 因子值越高，原子的可动性越大。通过检测 B 因子（在 PDB 文件中或者用 B 因子着色可显示的原子）可以推测某个原子是否可动。这是一种很好的方法，可以通过相互作用推测蛋白质的侧链是否稳定，同时可以用来研究在活性位点结合的配体哪一部分是稳定的，哪一部分是难于确定的。

除了 X 射线晶体衍射技术，核磁共振（NMR）波谱技术也能提供蛋白质及

蛋白质-配体复合物的原子解析结构信息。因为 NMR 实验是在溶液状态下进行的，因此不需要晶体。然而该技术对分子质量的限制（NMR 最大可以测量约 100kDa 大小的蛋白质，在有柔性连接臂的区域其大小也相应地受到限制）使其较难应用于确定蛋白质的结构。目前，只有一小部分的蛋白质结构通过 NMR 实验技术获得。由于计算机化学基因组学的研究主要倾向于选择晶体结构，因此，我们在此不详细阐述 NMR 技术如何对结构进行解析。但是，需要注意的是，有时 NMR 技术可以给出非常有价值的数据，因此，如果 NMR 数据可以获得的话，我们还是强烈建议使用 NMR 技术。

2. 理论蛋白质结构

实验所得的蛋白质结构信息可以很明确地作为计算机水平的研究起点。然而实验数据并不能时时获得（如 GPCR），因而产生了一些方法用来发展蛋白质以及蛋白质-配体理论上的结构。在缺失大量实验结构信息的情况下，化学基因组学就会应用理论蛋白质结构数据。

理论结构缺失一些实验方面的数据，如结构的解析、原子的 B 因子等，然而理论模型可以在短时间内建立和改进。理论结构是使用相关实验结构的信息，例如，建立一个蛋白质模型比较好的方法是使用一个已知实验结构的同源相关蛋白质的信息（见 2.2.2）。另一方面，从理论上模建配体结合形式（如 docking 计算）的过程也要用到某些蛋白质侧链与功能基团的相互作用的实验数据。

在有些时候，文献中没有相关的结构信息可用，那样，可以试着构建一个从头设计（*ab initio*）的蛋白质模型。通过能量方程的优化可以预测从头设计蛋白质模型，该类方程是用以描述氨基酸的物理学性质或者统计学优先性的。尽管经过了几十年的巨大努力，从同源的角度预测一个从头设计的三级结构还是十分困难的。从头设计的预测过程需要长时间的计算，并且预测结果通常是不可靠的。然而最近一些使用等级体系中的手段发展起来的方法似乎有希望产生低分辨率的结构[5]，即先建立局部结构，然后将其整合到整体结构中。

3. 蛋白质数据库

大体上有三种蛋白质数据库对于化学基因组学的计算机研究是必需的。序列数据库提供目标蛋白的同源信息；结构数据库提供目标蛋白和（或）序列相近的蛋白质的原子 3D 结构信息；其他的各种信息，如突变数据、活性数据等，也是很有价值的。

在序列数据库中可以很容易地找到蛋白质的氨基酸序列，如蛋白质数据库 SWISS-PROT（http://us.expasy.org/sport/）、TrEMBL（一个对 SWISS-PROT 数据库的计算机注解补充）[6]，以及 PIR（http://pir.georgetown.edu/）[7]。

SWISS-PROT是一个蛋白质序列数据库，力求提供高水平注解信息（如蛋白质的功能描述、区域结构、转录后修饰、变量等），避免冗余信息，与其他数据库高水平地整合。TrEMBL是对于SWISS-PROT数据库的计算机注解补充，包括所有未整合到SWISS-PROT的EMBL核酸序列条目的翻译。PIR是一项蛋白功能注解的公共资源，用以支持基因组/蛋白组学研究及科学发现。通过联合SWISS-PROT、TrEMB及PIR蛋白质数据库，同时联合蛋白质数据库工程（United Protein Databases，UniProt），可以创建一个蛋白质序列及功能的中央数据库[8]。

蛋白质及蛋白质-配体复合物的3D结构主要来源是RCSB蛋白质数据库（http://www.rcsb.org/pbd）[9]。这个数据库包括晶体结构、NMR结构，以及蛋白质、核酸、糖类及其复合物的结构。除了原子坐标，如果没有晶体结构还可以发现很多其他的补充数据，如溶解度、R因子、B因子和结晶条件等；而且蛋白质来源、结构问题、作者和引用都有提供。目前19 000多种的蛋白质及其复合物存放在该数据库，每一个都有一个4位数的ID密码，研究者可以在几个领域进行搜索，而且这些数据可以自由下载。这里需要提到的是结构的作者有一年的时间保留这个结构的发表，因此很少有数据是不可以用的。鉴于化学基因组学的计算机研究目的，这里有两种可能性。一种是寻找目标蛋白的构建好的结构，如果还没有构建好目标蛋白的3D结构，那就不得不寻找具有高度相似序列的蛋白质。这一点很容易做到，因为PDB可以帮助搜寻相似序列。当确定了一个或多个相似蛋白质的结构，需要通过同源模建的手段改进目标蛋白的理论结构。

RELIBASE+（http://relibase.rutgers.edu）是一个用来搜寻蛋白质-配体的数据库[10]，并且已发展成专门处理蛋白质-配体问题的数据库。RELIBASE+的特征包括配体结合位点有层理的详细分析、配体相似性、亚结构的搜寻以及蛋白质-配体和蛋白质-蛋白质相互作用的3D模式。RELIBASE+之后的数据库是PDB，该数据库包括来自NMR波谱技术或者X射线衍射技术的PDB形式的所有条目。一般来说不包括理论结构数据，只考虑模建一个配体的结构。在RELIBASE+中，一个结构中的非蛋白质残基都会被认为是配体。

虽然PDB数据库包括了一些理论蛋白质模型，但是MODBASE（http://guitar.rockefeller.edu/modbase）才真正拥有大量的理论模型，它包括多于800 000个可靠的模型及超过400 000种蛋白质区域的折叠形式[11]。MODBASE是一个注解比较蛋白质结构模型的相关数据库，这些模型是针对至少符合已知蛋白质结构的蛋白质序列的。

这里值得一提的是关于G蛋白偶联受体（GPCR）的数据库（http://www.gpcr.org/7tm）系统信息，因为它是药物产业的主要目标。这是一个搜集、整合、验证和传播GPCR上不同数据的分子类别特异（molecular-class-specific）的信

息系统。数据库主要依据序列、配体结合常数以及突变进行存储数据，同时也提供如序列比对、同源模建、质疑问题及可视化工具的计算性数据。

2.2.2 蛋白质的同源模建

1. 基于序列的队列

同源模建是一个多步骤过程，包括以下几项：

（1）认证同源蛋白。

（2）排列相同关系的序列。

（3）认证结构保守和结构可变的区域。

（4）运用已知结构的蛋白质作为模板构建保守区域（structurally conserved region，SCR）的3D结构。

（5）在结构可变区域（structurally variable region，SVR）内构建3D茎环结构。

（6）构建模型核心3D结构的侧链。

（7）评估及优化模型。

认证同源蛋白通常需要对蛋白质数据库运用不同的运算法则。

调整序列通常需要回答至少3个问题：①用哪个队列；②用哪个评分方法；③是否及怎样指定误差补偿（gap penalty）。BLAST[12]、FASTA[13]及Smith-Waterman[14]运算法则可以用来进行同源蛋白的认证。

可以建立一个20×20的矩阵来给队列打分，其中，已认证的蛋白质和那些与其具有相似特性的会比与其不同的蛋白质分值更高。打分方案分成以下几组。

- 专门考虑同一残基的基于特性的方案。
- 考虑互换氨基酸密码所需的DNA或RNA碱基改变数目，基于遗传密码的方案。
- 考虑理化性质的基于相似性的手段。
- 考虑序列队列中取代频率的基于取代的方案。

通过观察到的取代进行打分是序列队列很流行的方法，在此基础上发展了很多打分方案。

Dayhoff及其同伴们[15~17]在进行蛋白质演化分析的过程中发展了Dayhoff突变数据矩阵方法。矩阵可以就此演化的突变可能性给出在一个特定的演化时期一种氨基酸突变成另一种氨基酸的可能性，对PAM（percentage of acceptable point mutation）数值进行打分。一个低PAM的序列识别高同源序列的短队列，高PAM矩阵识别长的但是比较弱的队列。

在应用较少相关序列的局部多重队列来引入取代矩阵的过程中，Henikoff