第六章 蛋白质的功能域、结构及其药物设计

随着人类基因组全序列测定的完成，预示着基因组研究从结构基因组(Structural Genomics)进入了功能基因组(Functional Genomics)研究时代。研究基因组功能当然首先要研究基因表达的模式。当前研究这一问题可以基于核酸技术，也可以基于蛋白质技术，即直接研究基因的表达产物。测定一个有机体的基因组所表达的全部蛋白质的设想是由 Williams 于 1994 年正式提出的，而“蛋白质组”(proteome)一词是 Wilkins 于 1995 年首次提出。蛋白质组是指由一个细胞或组织的基因组所表达的全部相应的蛋白质。蛋白质组与基因组相对应，均是一个整体概念，但是两者又有根本的不同：一个有机体只有一个确定的基因组，组成该有机体的所有不同细胞都共享有一个基因组；但是，基因组内各个基因表达的条件、时间和部位等不同，因而它们的表达产物(蛋白质)也随条件、时间和部位的不同而有所不同。因此，蛋白质组又是一个动态的概念。由于以上原因，再加上由于基因剪接，蛋白质翻译后修饰和蛋白质剪接，基因遗传信息的表达规律更趋复杂，不再是经典的一个基因一个蛋白的对应关系，而是一个基因可以表达的蛋白质数目大于一。由此可见，蛋白质组研究是一项复杂而艰巨的任务。

蛋白质结构与功能的研究已有相当长的历史，由于其复杂性，对其结构与功能的预测不论是方法论还是基础理论方面均较复杂。统计学方法曾被成功地应用于蛋白质二级结构预测中，如Chou和Fasman提出的经验参数法便是最突出的例子。该方法统计分析了各种氨基酸的二级结构分布特征，得出相应参数(P_a,P_β和P_t)并用于预测。本章将简要介绍蛋白质结构与功能预测的生物信息学途径。

第一节 蛋白质功能预测

一、根据序列预测功能的一般过程

如果序列重叠群(contig)包含有蛋白质编码区，则接下来的分析任务是确定表达产物——蛋白质的功能。蛋白质的许多特性可直接从序列上分析获得，如疏水性，它可以用于预测序列是否跨膜螺旋(transmenbrane helix)或是前导序列(leader sequence)。但是，总的来说，我们根据序列预测蛋白质功能的唯一方法是通过数据库搜寻，比较该蛋白是否与已知功能的蛋白质相似。有 2 条主要途径可以进行上述的比较分析：

①比较未知蛋白序列与已知蛋白质序列的相似性；

②查找未知蛋白中是否包含与特定蛋白质家族或功能域有关的亚序列或保守区段。

图 6.1 给出了根据序列预测蛋白质功能的大致过程。由于涉及数条技术路线，所得出的分析结果并不会总是相一致。一般来说，数据库相似性搜索获得的结果最为可靠，而来自 PROSITE 的结果相对不可靠。

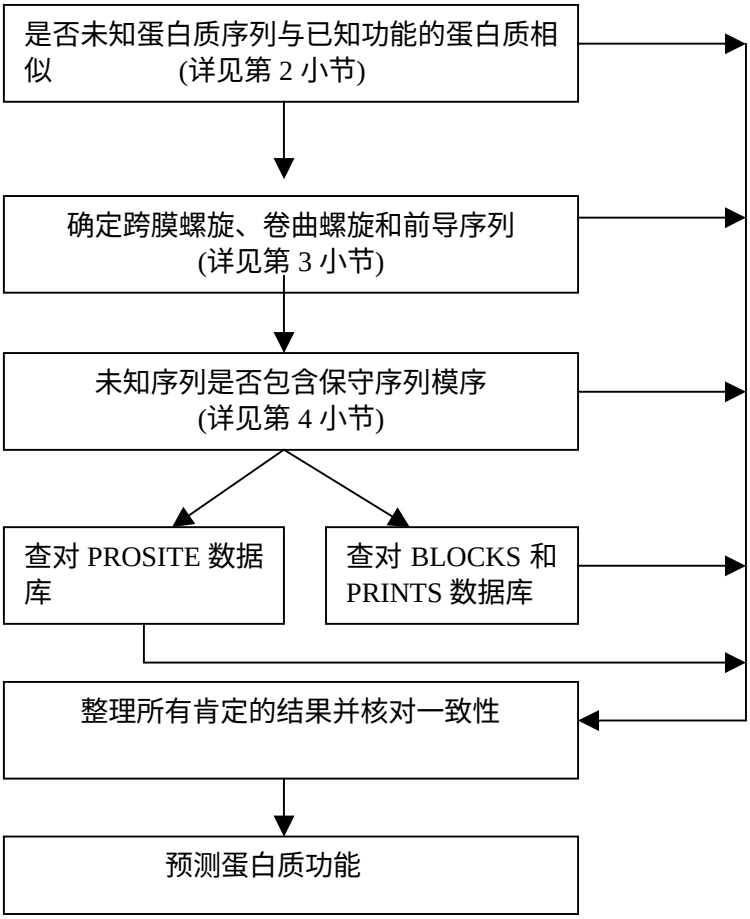


图 6.1 根据序列预测蛋白质功能的技术路线

二、通过比对数据库相似序列确定功能

具有相似序列的蛋白质具有相似的功能。因此，最可靠的确定蛋白质功能的方法是进行数据库的相似性搜索。具体的搜索方法可参见第三章，但应记住，一个显著的匹配应至少有 25%的相同序列和超过 80 个氨基酸的区段。

已有不少种类的数据库搜索工具，它们或者搜索速度慢，但灵敏；或者快速，但不灵敏。快速搜索工具(如BLASTP)很容易发现匹配良好的序列，所以没有必要再运行更花时的工具(如FASTA、BLITZ)；只有在诸如BLASTP不能发现显著的匹配序列时，这些工具才被使用。所以，一般的策略是首先进行BLAST检索，如果不能提供相关结果，运行FASTA；如果FASTA也不能得到有关蛋白质功能的线索，最后可选用完全根据Smith-Waterman算法设计的搜索程序，例如BLITZ(www.ebi.ac.uk/searches/blitz.html)。BLITZ不做近似估计(BLAST和FASTA根据Smith-Waterman算法做近似估计)，所以很花时，但非常灵敏。通常诸如BLITZ的程序能够发现超过几百个残基但序列相同比率低于 20～25%的匹配，这些匹配可能达到显著，但会被那些应用近似估计的程序错过。

还应注意计分矩阵(scoring matrix)的重要性。选用不同的计分矩阵有不少重要原因：首先，选用的矩阵必须与匹配水平相一致，例如，PAM250 应用于远距离匹配(<25%相同比率)，PAM40 应用于不很相近的蛋白质序列，而 BLOSUM62 是一个通用矩阵；第二，使用不同矩阵，可以发现始终出现的匹配序列，这是一条减少误差的办法。

除了选用不同的计分矩阵，同样可以考虑选用不同的数据库。通常可以使用的数据库是无冗余蛋白序列数据库 SWISS-PROT 和 PDB。其它一些数据库也可以试试，如可用 BLASTP 搜索复合蛋白质序列库 OWL (www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/owl_blast.html)。

二、序列特性：疏水性、跨膜螺旋等

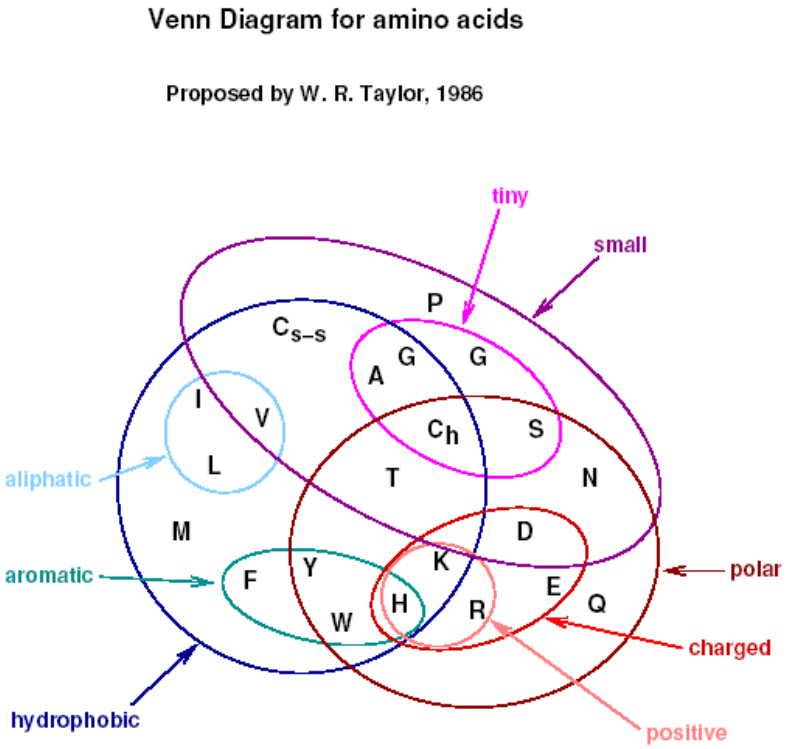
许多功能可直接从蛋白质序列预测出来。例如，疏水性信息可被用于跨膜螺旋的预测。还有不少小的模序(motif)是细胞用于特定细胞区室(cell compartment)蛋白质的定向。网上有大量数据资源帮助我们利用这些特性预测蛋白质功能。

疏水性信息可用 ExPASy(http://expasy.hcuge.ch/egibin/protscal.pl)的 ProtScale 程序创建并演示。这是一个很有用的工具，它能计算超过 50 种蛋白质的特性。程序的输入即可通过输入框将序列粘贴进去，也可输入 SWISS-PROT 的记录号。仅一项需要额外设定的参数是输入框的宽度，该参数将指示系统每次运行计算和显示的残基数，其缺省值为 9。如果想考虑跨膜螺旋特性，该参数设置应为 20，因为一个跨膜螺旋通常有 20 个氨基酸长度。图 6.2 是 ProtScal 程序的一个典型结果显示格式。

图 171 图 16.2

有多种方法可以预测序列的跨膜螺旋。最简单的方法是通过查找包含有 20 个疏水残基的区段，一些更复杂、更准确的算法不仅可以预测跨膜螺旋的位置，还能确定其在膜上的方向。这些方法都依赖于一系列已知跨膜螺旋特性的研究结果。TMbase 是一个自然发生的跨膜螺旋数据库(http://ulrec3.unil.ch/tmbase/TMBASE_doc.html)。相关的一些程序：TMPRED (http://ulrec3.unil.ch/software/TMPRED-form.html)、PHDhtm (www.embl_heidelberg.de/services/sander/predictprotein/predictprotein.html)、TMAP (http://www.embl-heidelberg.de/tmap/tmap/tmap_sin.html)和 MEMSAT (ftp.biochem.ucl.ac.uk)。这些程序将使用了不同的统计模型，总体上，预测准确率在 80～95%左右。跨膜螺旋是可以根据序列数据比较准确预测的蛋白质特性之一。

预测前导序列或特殊区室靶蛋白信号的程序：SignalP (http://www.cbs.dtu.dk/services/SignalP) 和 PSORT (http://psort.nibbac.jp/form.html)。另一个可从序列中确定的功能模序是卷曲(coil)螺旋。在这一结构中，二个螺旋由于疏水作用而缠绕在一起形成非常稳定的结构。相关的 2 个程序：COILS (http://ulrec3.unil.ch/software/COILS_form.html) 和 Paircoil (http://ostrich.lcs.mit.edu/cgi-bin/score)。



四、通过比对模序数据库等确定功能

经常会出现这样的情况：通过列线，未知蛋白质序列与数据库内已知功能的序列均相差较大，找不到可靠的匹配结果，相反，也许会发现与某一不知功能的序列相匹配。对于这一情况，仍然可以用生物信息学工具进行一些分析。

蛋白质不同区段的进化速率不同：蛋白质的一些部分必须保持一定的残基模式以保持蛋白质的功能，通过确定这些保守区域，有可能为蛋白质功能提供线索。例如，有许多短序列可以识别蛋白质活性位点或结合区域。整联蛋白(integrin)受体识别 RGD 或 LDV 配体模序(motif)，如果未知序列中包含有 RGD 模序，则可推测未知序列的一个功能可能是结合整联蛋白。这样的推测并不是说该蛋白质序列一定会结合整联蛋白(许多含有 RGD 的蛋白质并不结合整联蛋白)，但它的确为我们提供了一个可供试验的假设。还有些例子是保守序列位于酶活性位点、转录后修饰位点、协作因子结合位点或蛋白质分类信号等，不少有关这些保守模式(pattern)的生物信息学资源已经建立起来，并已用于在序列的搜索比对。

主要有二种方法可用于序列模序的查找。一种方法是查找匹配的一致(consensus)序列或模序。该技术的优点是快捷，模序数据库庞大且不断被扩充；缺点是有时不灵敏，因为只有与一致序列或模序完全匹配才会被列出，而近乎匹配的都将被忽略。这将使你进行更复杂的分析时受到严重限制。这时，第二种方法，一种更精细的序列分布型(profile)方法将发生作用。原则上，分布型搜索的是保守序列(不只是一致序列)，这样可以更灵敏地找出那些相关性较远的序列。但是分布型和分布型数据库的创建并非易事，它需要大量的计算和人力，因此，分布型数据库的记录数并没有模序数据库多。在实际分析时，应同时对这二种类型的数据库都进行搜索，其中在一个数据库中显著的匹配可能在另一个数据库中

被完全错过，反之亦然。

最知名的模序数据库是PROSITE(http://expasy.hcuge.ch/sprot/prosite.html)。PROSITE记录的典型形式（以酪蛋白激酶 II 磷酸化位点的一致序列为例）：[ST]-x(2)-[DE]，即一个丝氨酸(S)或酪氨酸(T)紧跟任意 2 个残基，然后再是一个D或E。另外记录中包含了位点其它一些重要信息，如位点的作用、在何处被发现等。

分布型(profile)数据库主要有 BLOCKS (http://www.blocks.fhcrc.org/blocks/)、PRINTS (http://www.biochem.ucl.ac.uk/bsm/dbbrowers/PRINTS/)和 ProDom (http://protein.toulouse.inra.fr/prodom/prodom.html)。正如其它生物信息学资源一样，这些数据库总是在规模和质量之间寻求平衡。对于分布型数据库的质量来说，还包括多序列列线产生的分布型。记录数最多的数据库是依赖于自动列线程序，得到的结果有时并非是最佳结果；而记录数少的数据库一般花很多时间用于分析，人工核对列线结果，力求产生高质量的结果。一般地，分析时应搜索所有的相关数据库，以保证没有任何的遗漏。BLOCKS 数据库是利用 PROSITE 数据库模序经无空位多序列列线构建而成，PRINTS 数据库(最小的数据库)的记录来自保守序列的多序列列线，而 ProDom 数据库(version33)数据则来自 9600 个蛋白功能区模序(domain motif)的列线结果。以上列出的数据库具体情况和输出结果(有时还挺复杂)等可参照各数据库的帮助说明。

第二节 蛋白质结构预测

一、蛋白质结构及其数据库

一般情况下，蛋白质的结构分为 4 个层次：

初级结构——蛋白质序列；

二级结构——α－螺旋和β－折叠片(β-sheets)模式；

三级结构——残基在空间的布局；

四级结构——蛋白质之间的互作。

近年来，另一个介于二级和三级结构之间的蛋白质结构层次——所谓蛋白质折叠(fold)已被证明非常有用。“ fold ”描述的是二级结构元素的混合组合方式。

根据序列或多序列列线预测蛋白质二级结构的技术已相对比较成熟(见下小节)，但三级结构的预测则相当困难。往往对于三级结构预测，只能通过与已知结构蛋白序列同源性比对来完成。已有不少相关数据库被建立起来用于蛋白质结构预测。这一方法已是目前进行三级结构预测的最准确方法(见第三小节)。但是这一方法并不总是奏效，因为大约有 80%的已知蛋白质序列找不到与之相似的已知结构的蛋白质序列。近年来，一些新方法被提出，这些方法可以不通过相似性比对来预测序列结构。

蛋白质结构数据库主要包括 PDB、NRL－3D、HSSP、SCOP 和 CATH 等，这些数据库的基本情况及网址请参阅第二章蛋白质数据库一节。

二、二级结构预测

已有大量有关根据序列预测蛋白质二级结构的文献资料，这些资料可大致分为二类：一是有关根据单一序列预测二级结构；二是有关根据多序列列线预测二级结构。

直到最近为止，二级结构预测才不被认为具有很高的随机性。大多数预测算法均是依据单一序列。即使是最著名的一些算法(如Chou-Fasman算法和GOR算法)也只有约 60%的预测准确率，而对于一些特定的结构，如那些富含 β－折叠片的结构，这些算法难以预测成功。预测失败的原因主要是单一序列所提供的信息只是残基的顺序而没有其空间分布的信息。两个方面的研究进展改变了这一状况：一是认识到多序列列线可被用于改进预测能力。多序列列线可被视为诱变遗传学试验中的自然突变状况，其对序列上单一位点变异的分析的确提供了该位点在蛋白质三级结构中的信息；二是神经网络已开始被用于根据序列预测结构。目前已有这样一个共识，即在有大量、高质量的多序列列线结果的情况下，蛋白质二级结构的预测将非常准确——通常准确率比以单一序列预测提高 10%。一些文献表明，一些程序(诸如 PHD)预测的准确率达到了目前最高水平。PHD(http://www.embl-heidelberg.de/predictprotein/predictprotein.html)提供了从二级结构预测到折叠(fold)识别等一系列功能。

三、三级结构预测

比对数据库中已知结构的序列是预测未知序列三级结构的主要方法。多种途径可进行以上这种比对。最容易是使用 BLASTP 程序比对 NRL－3D 或 SCOP 数据库中的序列。如果发现超过 100 个碱基长度且有远高于 40%序列相同率的匹配序列，则未知序列蛋白与该匹配序列蛋白将有非常相似的结构。在这种情况下，同源性建模(homology modeling)在预测该未知蛋白精细结构方面会发挥非常大的作用。在序列相同率为 25%～40%时，两条蛋白质将具有相同的折叠，但这时同源性建模将变得更加困难和不准确。

如果在比对 NRL－3D 数据库时没有发现匹配序列，接下去可试试 HSSP 数据库。这样做的一条最方便捷径是用 BLAST 或 FASTA 法搜索蛋白质序列库(如 SWISS-PROT、TREMBL 或 PIR)，然后利用诸如 SRS 等工具去检索任何超过 25%序列相同率的匹配序列，如果这些匹配序列在 HSSP 数据库中存在，则在该序列的注释(annotation)“ DR ”栏中将有说明(参见第三章)。如果未知蛋白质序列与某一 HSSP 数据库序列有明显大于 25%的序列相同率，则有把握地假定未知序列至少有与 HSSP 序列相同的蛋白质折叠模式。目前，NRL－3D 和 HSSP 数据库的记录数量可以保证 20%的蛋白质序列将找到已知结构的同源序列。

总的来说，同源性建模需要专业分子建模方法和分子图象资源的辅助才能进行。不妨到Swiss－Model网站(http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html)看看。Swiss－Model是一个蛋白质自动建模服务器，使用者可以直接发送一条序列或使用者自己完成的列线结果给该服务器用于同源性建模。

近年蛋白质结构研究的最主要进展之一，是有关“ 串线 ”(threading)算法和折叠识别。这些使人兴奋的技术可以在不存在已知结构同源蛋白质序列的情况下，预测所有可能的蛋白质结构。“ 这个未知蛋白序列会是什么结构呢？ ”我们也可以这样问：“ 我已经观察了已知结构蛋白质的各种折叠方式，未知序列是否会象这些已知结构中的某一个一样折叠呢？ ”第一个问题涉及几十亿种可能结构的搜索，而第二个问题涉及的是少于 1000 种结构的搜索。特定的蛋白质折叠被一而再，再而三地观察到——大部分新的经晶体衍射的蛋白将会与我们已知的折叠相关，这些过程使预测的成功机率不断提高。在串联算法中，未知序列以合适的方式被“ 串 ”到一个数据库某一折叠模板，然后计算该序列的能(energy)；在该序列与数据库中所有的折叠模板均“ 串 ”好后，可以进行计分比对，决定那些匹配达到了显著。折叠的识别技术目前还不是特别可靠的技术，只有在序列相同比率在 30%～50%时，

才有可能获得准确的估计。相关程序的结果也相当粗糙，大多数情况下难以作为同源性建模研究的依据。但是它是大多数蛋白质结构预测信息唯一可利用的工具。一些相关应用程序：TOPITS(http://www.embl-heidelberg.de/predictprotein/predictprotein.html)、frsvr(http://www.mbi.ucla.edu/people/frsvr/frsvr.html)、123D(http://www_lmmb.ncifcrf.gov/~nicka/123D.html)、THREADER和THREADER2(http://globin.bio.warwick.ac.uk/~jones/threader.html)和ProFIT(http:lore.came.sbg.ac.at/Extern/software/Profit/profit.html)。

第三节 计算机辅助药物设计[1]

开发一种新药需要平均10-12年，筛选1.5-2万种化合物，3-5亿美元。开发新药有两个瓶颈问题：疾病相关的靶标大分子的确定；具有生物活性的小分子药物的设计与发现。计算机辅助药物设计（computer-aided drug design，CADD）分为间接与直接设计，其基本原理“锁钥原理”：E. Fischer(1894)提出药物作用于体内特定部位，如同钥匙和锁的关系一样

间接药物设计

- 定量构效关系（quantitative structure-activity relationship，QSAR)：Hansch(1962)和Free & Wilson(1964)提出。不考虑化合物的空间结构，称为2D-QSAR。
- 3D- QSAR：CoMFA(比较分子力场分析)、距离几何（distance geometry)等
- 药效基因模型法

直接药物设计

- 以药物作用对象——靶标生物大分子的三维结构为基础,研究小分子与受体的相互作用，设计出从空间形状和化学特性两方面都可以很好与靶标分子“结合口袋”相匹配的药物分子。
- 分为全新药物设计（*de novo* drug design）和分子对接（docking）或数据库搜索两种方法。

全新药物设计

- 根据“结合口袋”的几何形状和化学特征设计药物分子
- 碎片连接法：基团或原子+适当的连接片段
- 碎片生长法：从靶标分子的结合空腔一端“延伸”出药物分子

分子对接（数据库搜索）

[1]本部分内容取自罗小民等，生物信息学与药物设计，见：赵国屏等主编，生物信息学，科学出版社，2001

⌗首先建立大量（几十到上百万）的化合物的三维数据库，然后用库中的分子与靶标分子进行“对接”（docking)，选出最佳构象的分子（前 50-100 个）供药理测验。
⌗ Kuntz(1982)发展了第一个 Dock 程序，这一方法取得巨大成功

设计实例：HIV-蛋白抑制剂