基于知识的蛋白质结构预测

赵善荣 唐 赟 陈凯先
(中国科学院上海药物研究所，上海 200031)

摘要 介绍了近几年基于知识的蛋白质三维结构预测方法及其进展．目前，基于知识的结构预测方法主要有两类，一类是同源蛋白模建，这种技术比较成熟，模建的结果可靠性比较高，但只适用于同源性比较高的目标序列的模建；另一类方法即蛋白质逆折叠技术，主要包括 3D profile 方法和基于势函数的方法，给出的是目标蛋白质的空间走向，它主要可用于序列同源性比较低的蛋白质的结构预测．

关键词 三维结构预测，同源蛋白模建，蛋白质逆折叠，序列结构匹配

基于知识的蛋白质结构预测（knowledge-based protein modeling）就是根据许多已知的蛋白质三维结构来预测序列已知而结构未知的蛋白质结构[1]．目前，随着分子生物学的发展和基因克隆技术的进步，许多蛋白质的序列已被测定出来，但其中许多序列的三维结构还没有用实验方法测定出来，结构数据的增加远慢于序列数目的增长，为了充分利用蛋白质的一级序列信息来研究结构与功能的关系，很有必要利用已知蛋白质的结构信息对这些蛋白质的结构进行预测．虽然蛋白质的序列繁多，但按其三维拓扑结构，可归属于许多不同的折叠类型[2]，这样的折叠类型估计有 500～700 个左右，在每一类中，各成员蛋白的三维结构类似，序列之间可能很相似，也可能不相似．若模建序列与模板序列经匹配（alignment）后的序列同源性（sequence identity）在 40% 以上，则它们的结构就可能属于同一折叠类型，可以用同源蛋白模建的方法预测其三维结构[3~5]．若匹配后的序列同源性较低（30% 以下），可用蛋白质逆折叠（inverse protein folding）技术把一个序列镶嵌入（threading）一结构已知蛋白质中，给出模建序列的拓扑结构．蛋白质的三维结构是其功能的基础，许多蛋白质（如酶）是药物作用的对象，根据蛋白质的三维结构可以进行药物作用机理和基于结构的药物分子设计研究，所以蛋白质的结构预测也受到药物分子设计研究者的广泛重视．本文主要对近几年同源蛋白模建技术和蛋白质逆折叠技术用于结构预测的研究进展作一综述．

1 蛋白质的同源蛋白模建

如果两个蛋白由同一种蛋白质分化而来，它们就具有相同或相近的功能与空间结构，因此，若知道了同源蛋白家族中某些蛋白质的结构，就可以来预测其他一些序列已知而结构未知的同源蛋白的结构．同源模建的序列之间同源性都较高，模建的整个过程可分为三步：模建序列（目标蛋白的序列）与模板序列的匹配；根据模板蛋白的三维结构构建目标蛋白的结构；对模建结构进行能量优化和分子动力学优化以及可信度评估．

1.1 模建序列与模板序列的匹配

同源蛋白模建首先需要把目标序列与序列库或结构库匹配，从库中按序列同源性和截断值挑选出一些同源性比较高的序列，然后把挑选出的序列与目标序列进行多重匹配．常用的序列库有 PIR，GENBANK，SWISS-PROT，EMBO 核苷酸序列库，结构库有 Brookhaven Protein Data Bank（PDB）等，搜寻软件主要有 FASTA 和 BLAST．

收稿日期：1995-11-17，修回日期：1996-02-26

1.2 根据模板结构模建目标蛋白结构

利用上步的多重匹配结果，确定模板结构等价位点套的初始集合，在初始集合的基础上，再旋转每一个模板的结构，使它们相互之间的Cα位置尽可能多地重叠在一起，不同模板中Cα在空间中若符合一定的重叠距离标准，它们相互之间的关系就是等价位点，许多这样的等价位点就构成了等价位点套．叠合结束后，即得到了同源蛋白的结构保守区（SCRs）以及相应的基架（framework）结构．模板结构匹配后，一般再用得到的同源体的SCRs的每一条序列与目标序列匹配，挑选并把目标序列上的高相似区定义为目标蛋白的SCR．

1.2.1 目标蛋白的结构保守区的主链模建：通过结构保守性的分析，可确定目标蛋白的结构保守区．在确定该区域的主链Cα原子坐标时，主要可采用有两种方法，一种是刚体装配法（rigid body assembly）[6]，即在同源蛋白族SCRs的相应各片段中，选取与目标蛋白SCR序列相似性最高的片段作为目标结构，另一种方法是加权平均法，即采用一种合适的加权方案，用构成基架的同源结构族的平均结构作为目标结构．权重的选取方法有多种，可以选用等权平均，也可取同源结构族与目标蛋白的序列同源性作为权，还可以给靠近基架的同源结构较大的权重，权重选取的方法对于模建的结构有一定的影响．近来，Srinivasan等[7]取同源性的平方作为权重来构建主链的结构，结果比较好．

1.2.2 结构变异区（SVRs）的主链模建：非保守区域主链结构较保守区域难预测，目前主要采用的有数据库查询和系统构象搜索方法，它们相互补充，各有优缺点，数据库查询方法快速方便，但并不能保证库中总有合适的片段可供选择，系统构象搜索方法则比较费时[8]．数据库查询方法的出发点很简单，即假定具备相似末端的等长片段，其空间结构类似．首先，在蛋白质结构数据库中，挑选分辨率比较高的结构数据，用它们的主链结构建立数据库．然后在非保守片段两端的保守片段上选取几个残基的Cα原子作为锚原子（anchor atoms），并计算两端对应的参考原子的两两距离．再在库中搜索序列长度满足该距离限制的等长片段，最后从搜索出的片段中挑选出一段，按锚原子的取向拷贝给待定结构片段．挑选片段对于变异区的结构模建十分重要，Topham和丁达夫等在这方面作了很好的工作，丁达夫等[9]首先依次把每一个片段插入到待模建的环区内，计算环区和其余保守核内的接触势能，并以此能量对每一个片段打分．Topham等[10]通过考察目标序列与周围环境的相容性对片段打分，这些因素包括主链的构象、溶剂可及性、氢键等等．一般来说，若变异区的序列长度小于7个残基，用数据库查询方法可得到较好的结果；如果长度在8个残基以上，由于查寻序列在库中出现的几率相对较小，往往搜索不到满足条件的片段．目前数据库查询方法已在许多软件包中得到广泛的应用，如Tripos公司的COMPOSER、Biosym公司的Homology、MSI公司的QUANTA/CHARMm等．

系统构象搜索的优点是不依赖于数据库，缺点是可能需要考虑的构象很多，并随着序列数目的增加，构象数目急剧上升．能量计算和结构数据的统计分析表明，氨基酸残基的二面角Φ和Ψ的分布主要集中在Ramachandran图中的几个区域内，为此，Moult等[11]对残基主链的二面角用12对Φ-Ψ的组合为代表来进行构象搜索，这样需要考虑的片段构象总数大大减少，但仍为12的n次幂（n为序列长度），实际应用时，可先对产生的构象初步筛选，主要判据有：端点位置是否合理和范德华表面碰撞．虽然作了大量简化，系统搜索方法计算量仍很大，因此也受到序列片段长度的限制．

如果变异区的序列长度较长，可以考虑采用Srinivasan发展的collar extension方法来解决变异区的主链模建问题[7]．最近，Sudarsanam[12]发展了一种基于Φ_i+1，Ψ_i二聚体数据库（Φ_i+1，Ψ_i dimer database）的

环区（loop）模建方法. 这种方法模建步骤是：先选择一些蛋白质构成数据库，然后统计序列上每对相连的氨基酸出现频率，把该频率记录在氨基酸对矩阵中（20×20），从该矩阵的每一个元素中又可以统计得出每对氨基酸对的 Φ_i+1，Ψ_i组合值分布，该矩阵就是 Φ_i+1，Ψ_i二聚体数据库. 最后，从变异区两端的保守区出发，对序列上相连的氨基酸对从库中任意指定一对 Φ_i+1，Ψ_i值，再对所产生的构象进行筛选，用这种方法大大减少了构象搜索的空间，该方法的成功之处是它考虑了主链结构中 Φ_i+1 和 Ψ_i的相关性.

1.2.3 侧链的结构预测：侧链的模建方法很多，但总的说来，绝大多数方法都是基于旋转构象库（rotamer library）的方法. 正如主链的二面角一样，侧链的二面角分布也不是在任意角度下都是等几率的，而是在某几个角度的可能性比较大，因此，我们可以根据已知的蛋白质结构，建立依赖于主链结构的 Φ，Ψ 值的侧链构象库，然后在此基础上，构建目标蛋白的侧链结构. 如果取与主链相连的构象库中的所有构象作排列组合，组合数目将会发生爆炸. Desmet 的死路消除法（dead-end elimination）用全局能量最低作为标准，能及时发现组合中死点，去除一些不可能的侧链构象，从而大大减少了组合的数目[13]. 近来，Koehl 等[14]发展的自洽平均场方法（self-consistent mean field）用于侧链的模建，结果很成功. 这种方法通过迭代优化构象矩阵 CM（i，j）来构建侧链，矩阵元素 CM（i，j）指与主链结构中第 i 个残基相连的第 j 个旋转构象出现的几率. 在优化第 i 残基的侧链构象时，先对其他残基的侧链用与之对应的矩阵行的几率平均构象来代替，然后依次计算第 i 行的每一个构象在平均场中的能量，据此再计算构象库新的几率分布. 重复迭代计算，直到构象矩阵 CM（i，j）最终收敛为止. 对矩阵行中每一行，取与最大几率相对应的构象作为侧链构象，从而完成了侧链的模建工作. 除了基于构象旋转库的方法外，还有许多其他的方法，比如 Hwang 等[15]最近发展一种基于神经网络的侧链结构的预测方法，结果就比较好.

近来出现了一种利用约束条件进行同源模建的方法，这种方法没有主链模建、环区模建、侧链模建等步骤，它直接根据序列匹配的结果，然后提取约束条件，并对目标序列模建出满足约束条件的模型来，这类方法类似于核磁共振中得到距离约束条件后，利用距离几何的方法计算出满足这些约束的结构. Havel 和 Sali 等对这种方法的发展作出了很大的贡献，其中 Havel 的方法[16]已由 Biosym 公司发展成商品化软件 Consensus. MODELLER 也是基于空间限制的一种模建软件[17]，该程序仅需输入目标序列和模板序列的匹配结果，程序自动产生限制条件并对目标序列进行模建，最后输出主链和侧链的重原子的三维结构模型.

1.3 对模建结构进行优化和评估

同源结构预测得到的蛋白质结构模型，通常含有一些不合理的原子间接触，需要对模型进行分子力学和分子动力学优化，消除模型中不合理的接触. 另外，模型中有些键长、键角和二面角也有可能不合理，同样也需要检查，现在已经有专门的软件完成这类工作，如 PROCHECK、PROSA Ⅱ 等. 许多用于蛋白质逆折叠的软件包本身也能对模建结构的合理性进行检查.

2 蛋白质的逆折叠技术与模建

严格说来，蛋白质逆折叠的含义是指已经知道一个蛋白质的三维结构，能否找到一个序列，使其折叠成该结构，若知道一个目标序列，问究竟什么样的三维结构与其对应，这是序列镶嵌问题，一般情况下，我们不严格区分这两个概念. 如果模建序列在序列库中没有搜索到与其同源性比较高的序列，难以用同源模建来预测其结构，这时，可以考虑采用序列镶嵌的方法来预测其三维拓扑结构. 目前，主要有两大类方法可用于蛋白质的序列镶嵌研究.

2.1 基于经验势函数的方法

这类方法的主要区别在于势函数（poten-

tial function）的选取不同，都希望通过对已知蛋白质进行结构分析，建立形式各异的势函数，从而用折叠构象的能量最低标准来指导目标蛋白质的序列镶嵌，看它能与哪一个已知的结构模板相匹配．在此，主要以 Bryant 的势函数[18]为例来对这类方法加以说明．

Bryant 的方法可分为三大步：第一步从 PDB 库中挑选出一些分辨率比较高的结构，然后测量每一种不同的氨基酸对（如 A-A, G-F等）之间的距离，并按间隔(0,5], (5,6], (6,7], (7,8], (8,9], (9,10]分类，然后用四元组（氨基酸 1、氨基酸 2、距离、频率）方式记录频率数据．第二步是把记录的数据转化成势函数，即把原四元组的记录转化成新的四元组（氨基酸 1、氨基酸 2、距离、势能）记录形式．假如把库中的每一个蛋白质的序列随机化，则每一个四元组的频率记录都有一个数学期望，根据这个期望频率，就可以把每一个记录频率转化成相对几率，再根据几率的 Boltzman 分布，把相对几率换算成势函数．第三步是把目标序列与每一种蛋白质的折叠纹基(folding motif）进行匹配，根据第二步建立的接触势函数计算构象的能量，并以能量为标准，寻找序列与纹基的最佳匹配方式．

Bryant 的方法所建立的势函数事实上是一种平均场势函数（potentials of mean force）．目前，具有知识的平均场势函数方法已渐渐开始成熟，并成为蛋白质结构研究中的一个重要工具[19]，利用它可以进行蛋白质的从头设计和蛋白质结构的从头预测．Kocher[20]对各种基于知识的势函数的结构预测能力及其影响因素作过评价．

近来，北京大学来鲁华等[21]利用蛋白质主链的极性分数及主链的二面角为参数，构建了一种基于蛋白质结构数据库的势函数，并把该势函数用于蛋白质的逆折叠研究中，发现该函数可成功地将蛋白质分子的天然构象从构建的构象库中识别出来．

2.2 基于剖面分析的方法——3D profile

3D profile 方法由 Einsenberg 小组的 Bowie 等[22]提出，并已经由 Biosym 公司发展成商品化软件 3D profile．这种方法首先根据溶剂可及性把残基分为三类：向溶剂暴露类（E）、部分埋藏类（P）、埋藏类（B），进而按照侧链面积中极性原子的暴露的百分数，把 P 分为 P1 和 P2，B 分为 B1、B2、B3，最后考虑残基的二级结构 α、β 和其他，这样就把残基的局部环境分成了 18 类。对已知结构蛋白库进行统计分析，计算各种残基在各类环境中出现的几率，建立 18×20 的矩阵（18 指 18 种环境类，20 代表 20 种氨基酸），矩阵元素 $S_{ij} = \ln(P(i,j)/P_i)$，其中 $P(i, j)$ 是残基 i 在环境 j 中出现的几率，而 P_i 是在任何环境中出现 i 的几率，这样我们就建立了每种氨基酸在不同环境类中的得分表．

当我们知道了一个蛋白质的三维结构时，通过它的原子坐标，可以计算出各残基的局部环境类，根据得分表，就把三维结构转化成一个 $N\times20$ 的剖面（N 为蛋白的长度），从而把整个结构库转化成一个 3D 剖面库．若已知一个序列，利用动态规划，把新序列与库中的每一个剖面比较，可寻找最佳匹配，对模建序列的空间结构进行预测．

Ouzounis 等[23]用接触矢量法描述蛋白质中的每一个残基的局部环境，然后在结构库中统计各种残基在各类接触中出现的频率，从而得到一个得分表，类似地，借助得分表，根据每个残基的局部环境，可把一已知结构转化成 $N\times20$ 剖面．Abagyan 等[24]运用溶剂化能发展了一种不依赖于数据库的统计信息来建立剖面的方法，首先将结构中的天然残基（native residue）突变成其他残基，计算残基溶剂化自由能，对结构中的每一个残基都进行突变和计算，结果就得到了一个 $N\times20$ 的剖面矩阵．

上面所述的同源模建和蛋白逆折叠两种模建方法在应用中是相辅相成的，并随着蛋白质结构数据的进一步积累和模建方法的进一步发展，模建结构的准确度也将会进一步提高．目前，比较同源模建已经是一门比较成熟的技术，当模建序列与模板序列的同源性在 40%

以上时，模建结构比较准确和可信．相对来说，用蛋白质的逆折叠技术来进行目标蛋白的结构预测还不是十分成熟，并且它只能判给出模建序列在空间的大致走向．由于目前我们很难从理论上来解决蛋白质的结构预测问题，因此更需要重视基于知识的蛋白质结构预测技术和方法的研究、发展，扩大这些方法的应用范围，进一步提高结构预测的准确度和可信度，更好地为深入研究蛋白质结构与功能的关系及基于结构的药物设计服务．

参 考 文 献

1 Johnson M S, Srinivasan N, Sowdhamini R *et al*. CRC Crit Rev Biochem Mol Biol, 1994; **29**: 1

2 Orengo C A, Jones D T, Thornton J M. Nature, 1994; **372**: 631

3 May A C W, Blundell T L. Curr Opin Biotech, 1994; **5**: 355

4 Sali A. Curr Opin Biotech, 1995; **6**: 437

5 Einsenhaber F, Persson B, Argos P. Crit Rev Biochem Mol Biol, 1995; **30**: 1

6 Kajihara A, Komooka H, Kamiya K *et al*. Protein Eng, 1993; **6**: 615

7 Srinivasan N, Blundell T L. Protein Eng, 1993; **6**: 501

8 Fidelis K, Stern P S, Bacon D *et al*. Protein Eng, 1994; **7**: 953

9 冯祖康，丁达夫．生物化学与生物物理学报，1995；**27**：173

10 Topham C M, Mcleod A, Eisenmenger F *et al*. J Mol Biol, 1993; **229**: 194

11 Moult J, James M N G. Proteins, 1986; **1**: 146

12 Sudarsanam S, DuBose R F, March C J *et al*. Protein Sci, 1995; **7**: 1412

13 Lasters I, Desmet J. Protein Eng, 1993; **6**: 717

14 Koehl P, Delarue M. Nature Struc Biol, 1995; **2**: 163

15 Hwang J-K, Liao W-F. Protein Eng, 1995; **8**: 363

16 Havel T F. Mol Simul, 1993; **10**: 175

17 Sali A, Potterton L, Yuan F *et al*. Proteins, 1995; **23**: 318

18 Bryant S H, Lawrence C E. Proteins, 1993; **16**: 92

19 Sippl M J. Curr Opin Struc Biol, 1995; **5**: 229

20 Kocher J P A, Rooman M J, Wodak S J. J Mol Biol, 1994; **235**: 1598

21 王彦力，来鲁华，韩玉真等．生物物理学报，1995；**11**：67

22 Bowie J U, Luthy R, Einsenberg D. Science, 1991; **253**: 164

23 Ouzounis C, Sander C, Scharf M *et al*. J Mol Biol, 1993; **232**: 805

24 Abagyan R, Frishman D, Argos P. Proteins, 1994; **19**: 132

Knowledge-based Protein Structure Prediction.

Zhao Shanrong, Tang Yun, Chen Kaixian (*Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai* 200031, China).

Abstract The methods and progress of knowledge-based protein three-dimensional structure prediction are mainly introduced. To date, the methods of knowledge-based structure prediction can be grouped into two categories. One is homologous protein modeling, which is relative mature and can give reliable results, but it is restricted to modeling of the target sequence that shares a high sequence identity with the homologous proteins. The other method, inverse protein folding, can be empolyed to predict the structure of protein with limited sequence homology, which mainly includes 3D Profile and potential-based functions, and gives the topology of the target protein.

Key words three-dimensional structure prediction, homologous protein modeling, inverse protein folding, sequence structure alignment