第1章 引 言

本章主要介绍本书的选题背景与意义，对领域内相关文献做出综述，同时简要说明本书的结构安排。

1.1 蛋白质结构预测概述

1.1.1 背景与意义

蛋白质是一类有机大分子，是组成生命体细胞、组织、器官等的最重要的成分之一，是生命活动的主要承担者，各种形式的生命活动几乎都有蛋白质的参与。因此，在当今生命科学研究的方方面面，蛋白质都占据极为重要的地位。

蛋白质构成的基本单元是氨基酸，氨基酸含有碱性的氨基、酸性的羧基及一个决定其种类的特异性侧链基团。多个氨基酸通过脱水缩合反应，彼此首尾相连组成一条肽链。之后，一条或者一条以上的肽链在空间中盘曲折叠，形成特定的空间结构，这便是蛋白质。目前，自然界中的氨基酸只有二十二种。根据它们的化学成分，蛋白质一定含有碳、氮、氧、氢元素，有些也含有硫、磷等元素。根据中心法则，蛋白质是生命信息流末端的最重要产物，其氨基酸排布序列对应着相应的基因编码。值得注意的是，蛋白质存在翻译后修饰过程，即其中一些氨基酸会发生受遗传信息调控的化学结构变化。

在漫长的生物化学探索中，人们逐渐意识到，蛋白质功能的行使往往需要特定的三维空间结构，即蛋白质的结构决定其功能，并由此拉开了结构生物学研究的大幕。通常来说，蛋白质的结构分四级。其中，一级结构指其氨基酸序列。二级结构则是对蛋白质上一个片段的主链原子的局部空间排列方式的描述(不考虑侧链原子的位置及此片段与其他片段的空间关系)，包括螺旋、片层及无规则卷曲等。蛋白质骨架的二级结构组成往往不是随机的，而是对特定结构及其功能存在高度特异性。三级结构指蛋白质中所有

原子的整体三维排列，二级结构之间通过几种弱相互作用(有时也会存在如二硫键之类的共价键等)保持其在三级结构中的位置。四级结构则存在于含有两个或两个以上独立亚基的蛋白质中，这些亚基在蛋白复合体中的排列即为该蛋白质的四级结构。

获取蛋白质结构是诸多生物学研究的第一步。掌握结构信息后，研究者才可能理解特定蛋白质行使其生理功能的原理，并以此为基础进行诸如小分子药物设计等的探究。那么，如何获得蛋白质的三维结构呢？传统的实验手段有 X 射线晶体衍射法(X-ray crystallography)、核磁共振法(nuclear magnetic resonance spectroscopy，NMR)及冷冻电镜法(cryo-electron microscopy，cryo-EM)等。根据 Berman 等(2000)对蛋白质结构数据库(protein data bank，PBD)的描述，目前每年解析得到的蛋白质结构在一万个左右，其中大多数是依靠 X 射线晶体衍射得到的。近年来(2013—2023 年)，随着冷冻电镜技术的普及(Wang et al.，2017a)及重构算法的革新(Li et al.，2013)，一些超大蛋白质复合物的高分辨率结构也逐渐可以被直接精确测定(Bai et al.，2015)。然而，在解析蛋白质结构的实验手段大获成功的同时必须指出其存在的缺点，那就是太过耗时耗力，即需要大量的人力及资金设备等的投入且实验周期较长。按照美国密歇根大学计算医学与生物信息系张阳教授的估算，实验解析一个较为复杂的蛋白质结构所需的花费为 250 000～500 000 美元。

蛋白质的序列信息决定了其空间结构。最重要的实验证明是 Christian Anfinsen 等在 20 世纪 50 年代关于蛋白质变复性的研究(Anfinsen et al.，1961；White，1961)。蛋白质在高温、极端 pH 值或变性试剂等的作用下丧失部分空间结构并且失去功能的过程称为蛋白质的变性，而其在恢复到自然的稳定条件下，天然结构和生物活性部分恢复的过程称为蛋白质的复性。在浓尿素溶液中，被纯化的核糖核酸酶 A 在还原剂存在下完全变性，其中，还原剂裂解四个二硫键，同时尿素破坏了维持蛋白结构稳定的疏水相互作用，因此核糖核酸酶的折叠结构整个崩塌，同时其催化活性完全丧失。当尿素和还原剂被去除时，Anfinsen 等(1963)又观察到，已经发生变性的随机卷曲结构又自发地重新折叠成有酶催化活性的正常生理结构。后续的很多研究都表明，蛋白质的氨基酸序列包含了其叠成天然三维结构所需的所有信息。

蛋白质序列信息的易得性奠定了从序列入手的蛋白质结构预测在当今生命科学研究中的重要意义。传统的多肽测序方法衍生出了一系列蛋白质

序列的测定技术，这些技术至今仍然在蛋白质化学中占有重要的地位。但是，目前来说，蛋白质的氨基酸序列大多是从基因组数据库中的 DNA 序列间接获得的。这就导致，随着基因测序技术的飞速发展，人类积累的蛋白质序列数据大规模增长，远远超过了相应蛋白质结构数据的积累，如图 1-1 所示。上文介绍的 Anfinsen 法则（Anfinsen，1973），也即蛋白质的氨基酸序列包含了其叠成天然三维结构所需的所有信息，又为从序列入手的蛋白质结构预测打下了理论基础。再加上其速度快、花费少、可以大规模并行计算等优点，蛋白质结构预测，尤其是从序列入手的蛋白质结构从头预测，在近些年来得到了充分的重视，吸引了大批优秀的研究者，也取得了令人欣喜的巨大进展（Baker and Sali，2001；Dill and MacCallum，2012）。

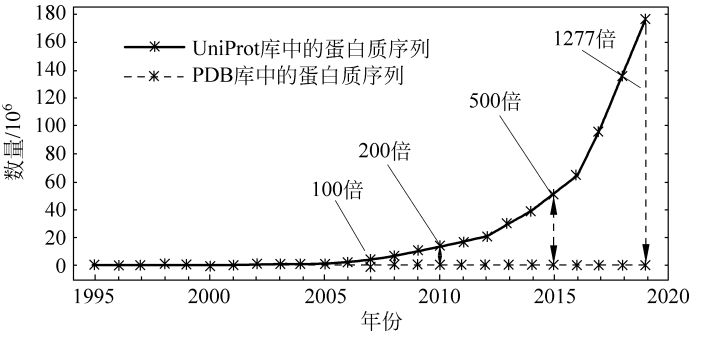

图 1-1　人类积累的蛋白质序列与结构数据的量随时间的变化

感谢美国密歇根大学计算医学与生物信息系张阳教授在其报告中对此图的慷慨分享。从图中可以看到，黑色的实线代表了在 UniProt 数据库中蛋白质的氨基酸序列随时间的变化趋势，呈爆发式、指数式的增长状态，与之相对的是在 PDB 中人类对蛋白质结构数据的积累变化（黑色的虚线），其基本呈平行于横轴的缓慢的线性增长。截至 2020 年，序列数据的量已达结构数据的 1277 倍

在蛋白质结构预测发展的过程中，国际竞赛 CASP（Critical Assessment of protein Structure Prediction）起到了重要的作用。CASP 是一个双盲性的比赛，组委会会在赛前向全世界的结构生物学实验室征集一些有望在最终评估参赛结果之前能够被解析，但是尚未被公开发表的蛋白质结构。在比赛过程中，所有参赛者能得到的只有这些目标蛋白质的序列信息，他们需要按照各自的方法在规定时间内完成预测并将预测结果提交至组委会。待所有题目都进行完毕后，组委会将根据一定的标准客观衡量所有参赛者的预测水平，之后组织学术研讨大会总结这个领域目前的情况及未来的发展方向（Kryshtafovych et al.，2019；Moult et al.，2016，2018）。CASP 比赛于

1994 年首次举办，之后每两年举办一次，为蛋白质结构预测提供了重要的评估平台，成为相关领域研究的算法测试基准。同时，也有一批非常有代表性的经典算法随着 CASP 比赛的进行而被研发及推广，如华盛顿大学 David Baker 教授团队开发的 Rosetta 系列(Das and Baker，2008)、密歇根大学张阳教授团队开发的 I-TASSER 系列(Roy et al.，2010；Yang et al.，2015；Zhang，2008)及芝加哥大学徐锦波教授团队开发的 RaptorX 系列(Kaellberg et al.，2012；2014；Xu，2019)等，会在下文中对这些方法进行较为详细的介绍。

作为相关领域的研究基础，蛋白质结构预测的飞速发展也带动了诸如蛋白质相互作用的预测(Sun et al.，2020)、蛋白质设计(Ben-Sasson et al.，2021；Cao et al.，2020)等方向的研究。随着越来越多不同背景、不同视野及不同技术特点的研究者的加入，蛋白质结构预测一定会取得长足的发展，产生更多影响深远的重大突破。

1.1.2 需要使用已有结构信息的预测方法

1.1.2.1 同源建模法

对于基因突变导致的蛋白质序列上氨基酸位点的突变，这一过程始终伴随着漫长的生物进化进程。我们在进化树中观察到的大多数突变是不影响蛋白质正常功能的，因为蛋白质功能发生变化的个体(尤其是与基本生命活动相关的关键性蛋白质，如呼吸作用中电子传递链上的蛋白质)将会面临更大的自然选择压力，往往被淘汰。同时，上文也介绍到蛋白质功能的行使必须以正常的三维结构为支撑，这也意味着一般留存下来的突变是不会剧烈影响蛋白质结构的。的确，研究者发现，同源蛋白质往往共享相似的结构，在进化过程中，蛋白质的三维结构比其氨基酸序列更加保守(Kaczanowski and Zielenkiewicz，2010)。

Chothia 等研究者在 1986 年就指出，当两条蛋白质的氨基酸序列拥有 20%以上的相似度时，它们对应的三维结构一般不会有巨大的差异(Chothia and Lesk，1986)，这便是蛋白质结构预测中同源建模(homology modeling)法的理论依据，即可以通过寻找结构已知的同源序列来估计目标序列的三维结构。

如图 1-2 所示，目前通用的同源建模法主要有以下步骤（Marti-Renom et al.，2000）。

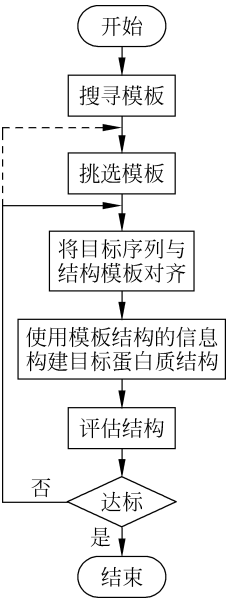

图 1-2　同源建模法的一般流程（Marti-Renom et al.，2000）

此图根据文献（Marti-Renom et al.，2000）绘制，展示了同源建模方法的一般流程。通常来说，同源建模法是一个迭代的过程。当产生的结构难以达到预测要求时，往往重复几个步骤，使其产生更多的结构，直到满足一定标准后停止

首先是在蛋白质结构数据库中寻找目标序列的同源模板序列。这一步常用 BLAST、PSI-BLAST 及 HHblits 等工具（Altschul et al.，1990；1997；Remmert et al.，2012）。合适的同源模板序列对后续操作及最终的结果表现都非常重要，因此需要对搜索结果做进一步筛选。常用的筛选标准为序列相似度、蛋白质功能相似度、表达调控相似度、二级结构相似度、相似覆盖范围与覆盖比率等。

其次是将筛选过的同源模板序列与目标序列进行对齐操作。在上一步的操作中，序列的高相似度区域已经有较好的对齐效果，在这一步，往往采用大尺度粗糙匹配算法（Rychlewski et al.，1998）以及随机搜索匹配算法（Mückstein et al.，2002）等提高序列中相似度较低区域的匹配准确度。

然后便是整个预测过程中最核心的部分，即构建目标序列的三维结构模型。这个过程得到的结果往往不是单一的，而是会产生一系列候选结构。

常用的方法大概可以分为两类：一类是先依靠同源模板的核心保守区域的结构搭建目标蛋白质的核心骨架，之后再通过对蛋白质片段库的搜索比对算法将剩余区域补齐(Wallner and Elofsson，2005)；另一类则是采用一种迭代优化的建模方式，具体来说就是根据找到的同源模板结构的统计信息构建目标序列对应结构的几何约束(常为概率分布的形式)，之后迭代地调整结构模型使其满足或者逼近这些几何约束(Šali and Blundell，1993)。

最后是质量评估与候选结构挑选。正如上文介绍的那样，模型构建这一步骤往往产生不止一个候选结构，需要通过一定的评估手段，如计算统计势能来比较其能量大小等，衡量这些候选结构在生物物理方面的合理性、是否符合天然肽的一般性质等条件，从中选择出最好的一个或者几个结构作为最终的预测结果。

在一般的同源建模蛋白质结构预测方法中，上述 4 个步骤往往需要循环迭代一次到多次。同源建模法对一般的蛋白质有非常好的预测效果，因为它们的同源模板序列可以在结构数据库中轻易地被找到，相关方法的预测精度也令人满意(Baker and Sali，2001)。但是对于同源序列稀少甚至无法找到的目标蛋白质，尤其是人为设计的蛋白质序列，其天然存在理论上的不足，这类方法往往对其束手无策。

著名的同源建模方法有 SWISS-MODEL(Schwede et al.，2003)，Modeller(Fiser et al.，2003)及 HHpred(Soding et al.，2005)等，这些方法为早期蛋白质结构预测的发展做出了重要的贡献，至今仍被广泛使用。

1.1.2.2 穿线法

根据上文中的介绍，我们知道，在进化过程中，蛋白质的结构相比蛋白质的氨基酸序列更加保守。那么，对于那些找不到同源序列的目标蛋白，是否还存在具有相似折叠模式的非同源结构模板呢？答案是肯定的。随着结构生物学研究的逐渐深入，人类积累了越来越多的蛋白质结构，将其按照结构相似性(也称蛋白质的折叠模式)归类，我们发现，折叠模式的增长近年来几乎停滞。最新的折叠模式数据库 SCOPe(structural classification of proteins extended-database)v2.07 显示(Chandonia et al.，2017)，目前所有的蛋白质结构仅可以分为 1232 个折叠模式。这就意味着，如果可以识别目标蛋白质的折叠模式，就可以将这种模式作为其骨架模板，对其结构做出较为准确的预测。

穿线法(threading)也称折叠模式识别法(fold recognition method)，就

是在这样的思路下应运而生的。那么如何确定目标蛋白质属于何种折叠模式并使用什么样的结构模板呢？目前的穿线法主要有两种技术路线：其一是依靠氨基酸序列对应的一维特征的匹配，这些特征主要有氨基酸的酸碱性、电性、亲疏水性、溶剂可及性、进化上的保守性及预测的二级结构归属等(Bowie et al.,1991)。其二是依靠预测的三维几何信息，如残基间距离等。目前使用最广泛的方法主要采用第二种技术路线，但是会采纳并融合第一种路线中的一维特征信息。其中，最具代表性的方法是密歇根大学张阳教授团队研发的 I-TASSER(Roy et al.,2010；Yang et al.,2015；Zhang,2008)，其他有名的方法还有 RaptorX(Källberg et al.,2012；2014；Peng et al.,2011)及 FALCON@home(Wang et al.,2016)等。值得注意的是，随着目前依赖模板的蛋白质结构预测的逐渐发展，穿线法和上文提到的同源建模法之间的界限越来越模糊，新的方法常常将它们混合使用以提高算法的预测性能。

1.1.2.3 片段组装法

尽管上文提到的穿线法已经对序列同源性的要求低了很多，但是还是需要对目标蛋白质的整体骨架进行折叠模式的确定与模板的搜寻。如果目标蛋白质的折叠模式不好确定，或者整体模板的搜索存在困难，是否存在将目标蛋白质化整为零，逐个片段搜寻相关的结构模版，之后再从零到整，将这些片段组装起来的方法呢？

Bowie 等在 1994 年提出了应用 9 个残基长度的短片段进行结构模板搜索，并最终拼接为完整的目标蛋白质结构的方法(Bowie et al.,1994)，这就是后来被广泛应用的片段组装法(fragment assembly method)的雏形。片段组装法一般包括片段库构建，片段结构模板搜索(也称构象搜索)与基于势能函数的片段选择、组装与最终结构的调整等部分。华盛顿大学的 David Baker 教授团队研发的 Rosetta 程序(Simons et al.,1997)是这个方向影响力最大、最被广泛使用的程序。Rosetta 程序吸引了很多研究者，形成了一个良性的开发和维护社区，衍生了许多不同的功能，且其源代码有较为详细的注释与使用说明(Khare et al.,2015)。下面就以 Rosetta 程序为例，对片段组装法的步骤进行简要说明。

Rosetta 程序官方的片段库构建方法为 NNMake(Gront et al.,2011)，其中包含长度分别为 9 和 3 的片段。对目标蛋白质的每一个位置来说，NNMake 都会使用诸如序列位点的相似性、预测的二级结构归属、残基间

的几何信息等作为打分依据，在结构数据库中选择 200 个 3 氨基酸片段及 200 个 9 氨基酸片段，组成这个位置的片段库。

在构象搜索方面，Rosetta 程序先使用片段库中的片段（包括 9 长度和 3 长度片段）对目标蛋白质相应位置上的氨基酸片段进行结构替换（置换相应的二面角），之后，在基于先验知识的势能函数的加持下进行蒙特卡罗模拟以将片段组装，这个过程中侧链一般使用质心表示。最后，将侧链添加回来，使用基于物理的势能函数对全原子模型进行局部微调以得到最终的预测结果。

1.1.3 完全基于序列信息的预测方法

1.1.3.1 基于序列的蛋白质残基间几何信息的预测

蛋白质残基间的相互作用使其形成了一定的空间结构。除 PDB 文件中那样直接使用原子的空间坐标来描述蛋白质的结构之外，还有很多其他的结构描述方式，如二面角序列、残基间距离矩阵等。其中，残基间距离矩阵与相应的蛋白质结构具有很好的一一对应关系，即拿到一个蛋白质的全原子坐标就可以轻松地求出任意一对残基的空间欧几里得距离，进而构建距离矩阵。同样地，如果有残基间距离矩阵，也可以根据多维尺度（multi-dimensional scaling）算法对其空间结构的原子坐标进行推导。因其优良的性质，诸如平移旋转不变性（蛋白质结构在空间中发生平移或者旋转，虽然原子坐标会发生变化，但是其残基间距离矩阵不发生变化）等，残基间距离矩阵得到了广泛的研究与应用。

一个氨基酸残基由多个原子组成，在计算残基间距离的时候，用什么样的坐标来代表这个残基呢？同时，在算法及算力水平还不够高的时候，直接预测、处理并使用蛋白质的残基间距离有一定困难，能否在解决问题的大前提下，对残基间距离做进一步的抽象呢？对于第一个问题，不同的研究者选择了不同的残基间距离的定义，其中有重原子最小距离（Berrera et al.，2003）、全原子最小距离（Mirny et al.，1996），C_α 距离（Vendruscolo et al.，1997）与 C_β 距离（Fariselli et al.，2001）等。对于第二个问题，研究者们在选定适当的距离阈值之后，常常通过实际距离大于或小于这个距离阈值来将残基间距离转换为 0 或者 1 的距离标签，这个标签称为残基接触（contact）。

根据 CASP 比赛（Schaarschmidt et al.，2018a），残基接触的一般定义为：若一对残基的 C_β 原子间的欧式距离小于 8 Å，其存在接触，反之则不

存在接触。对于没有 C_β 原子的甘氨酸残基，使用其 C_α 原子代替定义中的 C_β 原子。从这个定义来看，残基接触矩阵是稀疏的。根据统计，平均每个残基大约和 7 个其他残基发生接触。这一优良的性质对之后的预测结果处理及应用上会有很大帮助。在得到蛋白质的残基接触矩阵后，可以根据含有约束优化的一般算法对其结构进行方便的预测（Vassura et al.，2008；Vendruscolo et al.，1997）。因此，蛋白质残基接触预测成了蛋白质结构预测领域的重要研究方向。

目前来说，基于多序列比对（multiple sequence alignment，MSA）的共进化信息（co-evolutionary information）提取是蛋白质残基接触预测方法的主要基础。共进化是指空间联系紧密的氨基酸残基往往会出现同时突变，其存在高度的隐含相关性。这个想法非常朴素，下面举例说明。若一对残基带相反的电荷，则它们的接触可以通过静电相互作用稳定蛋白质的结构。此时，若其中一个残基由正电氨基酸突变为负电氨基酸，则这一对残基同时带有相同的电性，在能量的作用下它们不能再保持接触的状态，于是蛋白质的局部结构遭到破坏，甚至其功能也会受到影响，对应的生物个体会受到极大的自然选择压力，进而有可能被淘汰。因此，在长期的自然选择压力下，这些空间联系紧密的氨基酸残基往往会同时出现相关的突变以维持对应蛋白质的正常结构和功能。

正如上文中介绍的那样，目前人类积累的蛋白质序列数据仍然呈爆发式的增长，因此通过算法能找到的目标蛋白质的多序列比对结果所包含的信息量也会越来越巨大，导致通过残基接触预测的蛋白质结构预测手段拥有巨大潜力。近年来，随着深度学习技术在这个领域的成功应用，涌现了一大批优秀的残基预测算法，如芝加哥大学徐锦波教授团队开发的 RaptorX-Contact（Wang S et al.，2017b；Xu，2019）、格里菲斯大学周耀旗教授团队开发的 SPOT-Contact（Hanson et al.，2018）及我们实验室开发的 AmoebaContact（Mao et al.，2020）等。

RaptorX-Contact 是一种非常典型的蛋白质残基接触预测方法，其架构如图 1-3 所示。它首次将计算机视觉领域内大获成功的深度学习网络架构——残差网络（residual network，ResNet）引入蛋白质残基接触预测。RaptorX-Contact 先使用一个一维（ID）的残差网络得到一维特征，将其转化为二维之后，和共进化信息等二维特征连接，再使用二维（2D）的残差网络接收、处理与整合输入信息，最终得到对目标蛋白质残基接触图谱的预测。SPOT-Contact 则是在残差网络的基础上考虑到蛋白质序列信息的

有序性，结合使用了循环神经网络（具体为长短期记忆网络，LSTM）。AmoebaContact 除使用网络架构自动搜索技术对残基接触预测这一任务产生最合适的深度网络架构外，还创新性地打破了以往仅使用 8 Å 作为残基接触判断的距离阈值，通过引入多个不同的阈值对目标蛋白质的结构信息进行更好的描述。

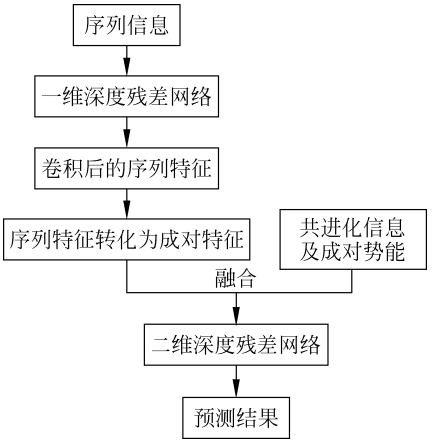

图 1-3 RaptorX-Contact 的网络架构（Wang et al.，2017b）

此图的绘制参考了文献（Wang et al.，2017b），简要示意了 RaptorX-Contact 的网络架构，由两部分残差网络（ResNet）组成，分别为一维残差神经网络和二维残差神经网络

在 2018 年的 CASP13 比赛中，Google 公司的 DeepMind 团队开发的 AlphaFold（Senior et al.，2020）大放异彩，夺得头筹。在 AlphaFold 的诸多创新点中，从对蛋白质残基接触的预测转向对其残基间距离的预测绝对是重要的变革之一。残基间的距离相比二值的残基接触，带有更加详细、精确的结构信息，对后续蛋白质折叠的指导会有更加重要的意义。与 AlphaFold 几乎同时或者在其发布后不久，这个领域的主流方法就转向了蛋白质残基间距离的预测，涌现了一批经典的算法，如升级后的 RaptorX-Contact（Xu，2019）、trRosetta（Yang et al.，2020）及 CopulaNet（Ju et al.，2020）等。

其中，trRosetta 因为引入了残基间转角的预测作为距离预测的补充，同时使用了方便的能量最小化函数 MinMover 来梯度下降地优化预测的蛋白质结构而被广泛讨论和使用。这里以 trRosetta 为例简单介绍蛋白质残基间几何信息的预测及其应用。从目标序列的 MSA 中抽提的输入特征进入一个有 60 个结构模块的深度残差网络。这个网络是多任务的，因为无论