September 25, 2014, 30(9): 1362−1371
©2014 Chin J Biotech, All rights reserved

综 述

预测蛋白质可结晶性研究进展

周仁斌，卢慧薏，尹大川

西北工业大学生命学院 空间生物实验模拟技术国防重点学科实验室，陕西 西安 710072

周仁斌, 卢慧薏, 尹大川. 预测蛋白质可结晶性研究进展. 生物工程学报, 2014, 30(9): 1362−1371.
Zhou RB, Lu HM, Yin DC. Progresses in predicting the crystallizability of proteins. Chin J Biotech, 2014, 30(9): 1362−1371.

摘 要：解析蛋白质的三维结构具有重要的生物学意义，更是蛋白质功能研究和理性药物设计的基础。目前解析蛋白质结构最重要的方法是 X-射线衍射晶体学解析技术。但是运用该技术解析蛋白质结构的关键是获得高质量的蛋白质晶体。然而，据统计仅有 42%的可溶纯化蛋白质能够得到晶体，即不同蛋白质的可结晶性表现不同。由于实验方法验证蛋白质的可结晶性耗时耗力，因此，有研究者运用计算机模拟的方法预测蛋白质的可结晶性，从而节省资源与成本并且提高实验的成功率。本文结合我们的研究工作，介绍了几种目前较为成功的蛋白质可结晶性预测方法及其研究途径。

关键词：蛋白质结构，X-射线衍射晶体学，蛋白质可结晶性预测

Progresses in predicting the crystallizability of proteins

Renbin Zhou, Huimeng Lu, and Dachuan Yin

Key Laboratory for Space Bioscience & Biotechnology, *School of Life Sciences*, *Northwestern Polytechnical University*, *Xi'an* 710072, *Shaanxi*, *China*

Abstract: Determination of protein 3-dimensional structure offers very important information in biology researches, especially for understanding protein functions and redundant drug design. The X-ray crystallography is still the main technique for protein structure determination. Obtaining protein crystals is an essential procedure after protein purification in this technique. However, there is only 42% of soluble purified proteins yield crystals by statistics. Experimental verification of protein crystallizability is relatively expensive and time-consuming. Thus it is desired to predict the protein

Received: December 17, 2013; **Accepted:** April 1, 2014
Supported by: National Natural Science Foundation of China (No. 31170816), National Basic Research Program of China (973 Program) (No. 2011CB710905), Fundamental Research Foundation of Northwestern Polytechnical University in China (No. JC20110285).
Corresponding author: Dachuan Yin. Tel/Fax: +86-29-88460254; E-mail: yindc@nwpu.edu.cn
国家自然科学基金 (No. 31170816)，国家重点基础研究发展计划 (973 计划) (No. 2011CB710905)，西北工业大学基础研究基金 (No. JC20110285) 资助。
网络出版时间：2014-05-13 网络出版地址：http://www.cnki.net/kcms/doi/10.13345/j.cjb.130639.html

crystallizability by a computational method before starting the experiment. In this paper, combined with our own efforts, some successful *in silico* methods to predict the protein crystallizability are reviewed.

Keywords: protein structure, X-ray crystallography, protein crystallizability

目前，蛋白质晶体学越来越受到广大研究者的青睐。因为蛋白质的结构解析是其功能研究[1]、疾病治疗[2]和药物设计[3-4]等的基础。在解析蛋白质结构的两种主要方法 (X-射线晶体学衍射技术[5]和核磁共振技术[6]) 中，X-射线衍射技术 (X-ray diffraction measurement，XRD) 是蛋白质结构解析的主要手段。到目前为止，PDB 数据库中 88.4%的蛋白质结构是用 XRD 方法获得的 (数据统计截止 2013 年 12 月)。用 XRD 方法确定蛋白质三维结构的实验过程包含多个环节，在各个阶段都有较高的失败率，这就会增加蛋白质结构获得的平均花费[7]。其中两个主要的步骤是蛋白质的纯化和结晶。据统计，仅仅只有 42%的可溶纯化蛋白质能够得到晶体[8]。因此，研究者提出影响蛋白质结晶成功率的内在因素是蛋白质本身所具有的可结晶性[9]。如果有一种预测方法，有允许的准确度，可以预测某个蛋白质的可结晶性，那么就可以显著地降低蛋白质结构解析的成本并提高成功率。因此，在开展结晶实验之前，基于蛋白质的序列及其特性，用计算机模拟的方法来预测蛋白质的可结晶性是非常必要的。随着网络数据库的不断完善与各种生物信息学软件的不断发展，我们可以方便地利用计算机模拟技术处理蛋白质结晶的相关问题。本文将对目前已有的预测蛋白质可结晶性的方法进行分析和总结。

1 蛋白质可结晶性预测方法的发展

首次基于蛋白质氨基酸组成来预测蛋白质可结晶性的方法是由 Smialowski 及其同事提出的 SECRET 方法[10]。他们发现蛋白质序列中氨基酸的组成和蛋白质的结构密切相关。他们首先从 PDB 数据库[11]中提取出序列长度为 30−200 个氨基酸的蛋白质的序列信息。以氨基酸的序列和氨基酸的疏水性为特征，并挑选出与蛋白质结晶密切相关的 12 个氨基酸 (R、N、D、Q、E、H、L、F、S、T、W 和 V) 来构建预测模型，在文中作者称为这些氨基酸为单字符 (One-word size) 特征。并且假设如果一个蛋白质序列中这些氨基酸的比重比较大，那么该蛋白质结晶的可能性就比较高。另外他们假设用 XRD 方法解析其结构的蛋白质是可以结晶的，而采用核磁共振方法解析其结构的蛋白质是无法结晶或难以结晶的。这种假设的基础是核磁共振方法解析蛋白质结构时价格昂贵，周期长，分辨率较低。而 XRD 是首选的蛋白质结构解析方法。虽然，后一种假设是值得怀疑的，但是 PDB 数据库中缺乏不可结晶蛋白质的相关信息，所以这种假设不可避免，也具有一定的可靠性。该假设是 SECRET 方法的基础，因为需要有正数据集 (可结晶的蛋白质) 和负数据集 (不可结晶的蛋白质) 对预测结果进行评价。SECRET 的分类方法采用的是 SVM 作为一级分类器[12]，贝叶斯网络 (Bayesian network) 作为多元分类器[13]。用十倍的交叉验证试验表明，SECRET 的预测准确率为 66.9%。这个结果尽管不是很理想，但是它是用计算机模拟方法预测蛋白质可结晶性的首次尝试。

2007 年，Chen 等提出了另外一个预测模型——CRYSTALP[8]。他们假定蛋白质的可结晶性和氨基酸的碱基对组成相关。和 SECRET 方法的假设一样，如果蛋白质序列中这些碱基对的含量比较高，那么该蛋白质的可结晶性就越高。文中还考虑了一个碱基对之间被 0、1、2、3 和 4 个氨基酸隔开的情况。这样的话，构建模型时输入的特征就比较多，不利于建模。所以 CRYSTALP 方法运用相关特征选择方法和贪婪算法[14]在 2 020 (400(4 + 1) + 20 = 2 020) 个氨基酸碱基对中挑选出 45 个与结晶最相关的氨基酸对，作者把这些特征称为双字符 (Two-word size) 特征。CRYSTALP 的数据来源也使用了 SECRET 所使用的 PDB 数据库。最终，CRYSTALP 的预测准确率为 77.5%，该结果优于 SECRET 方法。但是 CRYSTALP 的特异性仅仅为 71.3%，主要原因是该方法没有处理好正负数据集不平衡的问题，其中负的数据集 (不可结晶的蛋白质) 比正的数据集 (能够结晶的数据集) 要少的多。

除了蛋白质序列的氨基酸组成，其他研究者也注重用氨基酸的物理化学特性来预测蛋白质的可结晶性。氨基酸的这些特性对蛋白质的可结晶性有很大的影响。因此，2006 年，Overton 和 Barton 提出了另外一个预测方法——OB-Score[15]。该方法主要是对目标蛋白质的可结晶性进行排序。它是以蛋白质的等电点 (PI 值) 和总体平均疏水性 (Grand average of hydropathy, GRAVY 值) 作为预测特征。用 R 软件包[16]来处理这些特征，得到预测的模型。OB-Score 可以用来比较目标蛋白质的 PI-GRAVY 和先前已结晶的蛋白质之间的相似性。对目标蛋白质所计算出的 OB-Score 值越高，则表明该蛋白质从克隆、表达、纯化到结构确定各个过程的成功率越大。OB-Score 对 PfamA 家族[17]的蛋白质的预测准确率为 73.4%。因此 OB-Score 可以用来挑选出更容易成功结晶的蛋白质。但是该方法没有给出具体的分类准则。

2007 年 Slabinski 等又研发了一种基于网络的预测系统——XtalPred[18]。该方法一次性最多可以运行 10 个蛋白质序列，另外它对蛋白质的序列长度限制扩大到 50−1 000。该预测方法主要是把各个独立的预测特征整合成单个的可结晶性评分。根据评分可以把结晶可能性分为 5 类：最优、次优、平均、困难和非常困难。XtalPred 运用了几个在线的生物信息学软件来计算蛋白质结晶可能性的评分。但是该方法的预测准确率还没有被证实，因为某些在线的生物学工具无法给出某个特征的准确值。

2008 年，Overton 提议了另一种蛋白质可结晶性预测方法——ParCrys[19]，一种 Parzen 法来评估蛋白质产生衍射质量晶体的特性。PDB 数据库提供训练集，同时 TargetDB[20]和 PepcDB[21]数据集用来确定特征选择数据集和测试数据集。结果表明 ParCrys 法优于 OB-Score、SECRET 和 CRYSTALP 法，该方法的准确率和 MCC 的值分别为 79.1%和 0.582。

2009 年，研究者在 CRYSTALP 的基础上发展出新的预测方法 CRYSTALP2[22]。它是利用蛋白质氨基酸的组成和氨基酸的组合，等电点和疏水性来预测蛋白质可结晶性的。预测的蛋白质序列的大小不受限制，预测的准确率也进一步提高。结果表明，CRYSTALP2 的预测准确率、MCC 和 AROC 比 CRYSTALP、OB-Score、SECRET 方法高，和 ParCrys 与 XtalPred 方法接近。说明 CRYSTALP2 也是一种较好的预测蛋白质可结晶性的方法。

2010 年，Kandaswamy 等提出另外一个预测方法 SVMCRYS[23]。该方法综合考虑了影响蛋白质结晶的各种因素特征，最后用著名的数据最小化软件包 WEKA[24]选择出最优的特征。SVMCRYS 方法是利用支持向量机 SVM 把蛋白质分为两类：能够结晶 (Amenable to crystallization) 和难以结晶 (Resistant to crystallization)。SVMCRYS 方法对蛋白质序列的大小没有限制，用了几个具有代表性的特征来预测蛋白质的可结晶性，与其他预测方法相比，具有较高的准确率。

如何选择最相关的特征集是预测准确率的关键。因此 2013 年，Hsieh 等[25]首先收集了前面几个预测方法中所使用的蛋白质一级结构的 74 个特征，采用了 F-score 和信心增益 (Information gain, IG) 两个特征选择模型[26]来选择出与蛋白质结晶最相关的 48 个特征。这 48 个特征用 AdaBoost[27]来建模，此外也用 SVM 作为对比的预测模型。TargetDB 数据库评估这两种分类模型的实验结果表明五倍交叉验证的准确率为 93%，敏感性为 95.5%，特异性为 86.1%。但是该方法只考虑了来源于蛋白质序列中的特征，实际上蛋白质纯化和结晶的条件也影响 X-衍射的结果。由于 TargetDB 数据库的限制，这些特征没有考虑。

除此之外，基于蛋白质的序列和物理化学特征还发展了其他许多的计算机模拟方法 (PPCinter[28]、FRCRYS[29]、CRYSpred[30]、PPCpred[31]、HyXG-1[32]等)。而且上述的某些预测方法已经成功应用于结晶实验中。例如 OB-Score 和 XtalPred 方法已经成功应用于结构基因组学 (Structural genomics, SG) 对于目标蛋白的选择上[30]。特别是 XtalPred 方法，已经有多篇文章报道了其应用。Dom Bellini 应用 XtalPred 预测方法成功解析出了酶激活 HD-GYP 结构域 (Enzymatically active HD-GYP domain) 蛋白的结构[33]；Gómez-García 通过 XtalPred 方法分析出 cyclin M2 (CNNM2) 的 $CNNM2_{429–584}$，$CNNM2_{429–589}$ 片段的结晶可能性很大，从而成功纯化结晶获得两个片段的晶体，获得初步的衍射数据[34]；Abhinav Kumar 在研究核因子相关 K-B 结合蛋白 (Nuclear factor related to kappa-B-binding protein, NFRKB) 时，由于全长的蛋白无序区域太多，难以获得晶体。所以他们利用 XtalPred 方法预测出了 16 个包含目的结构域而且比较容易结晶的片段，最终其中一个片段成功得到高质量的晶体，从而解析出其结构[35]。

在上述各种预测蛋白质可结晶性方法思路的基础上，本实验室也提出了一种预测蛋白质结晶沉淀剂的方法。首先我们通过数据挖掘，发现蛋白质序列相似性和蛋白质结晶沉淀剂之间存在显著的相关性[36]。有了这个理论基础，我们从 PDB 数据库中提取出所有用 X-射线衍射方法解析出结构的蛋白质的序列信息和结晶沉淀剂信息，按照蛋白质序列相似性和结晶沉淀剂之间的相关性构建了预测模型，最后通过严格验证，我们该模型的预测准确率为 66.7%。而且我们还构建了网络服务器来实现这一方法的在线使用。这充分说明了我们可以尝试用计算机模拟的方法来解决蛋白质结构解析过程中遇到的各种难题。

总之用计算机预测蛋白质结晶相关问题时，一般分为 3 个步骤：1) 获取数据 (训练数据集、测试数据集和验证数据集)，并确定预测特征；2) 构建预测模型；3) 评估构建模型的性能。后文将对这 3 个步骤所涉及的具体问题逐

一介绍。

2 数据的获取

对于计算机模拟方法和数据挖掘工作而言，数据的来源至关重要。为了解决用 XRD 方法来解析蛋白质结构过程中的难题，早期采取的措施是建立一个完整的数据库，该数据库既要包含成功解析出结构的蛋白质的信息，也要包含尝试后失败的蛋白质信息。该思想是 2000 年 Stevens 提出的[37]。2003 年，Rodrigues 和 Hubbard 又提出，随着结构基因组学的发展，有价值的实验数据不断积累，研究者可以利用这些数据来改进基于氨基酸序列而构建的相关预测方法[38]。

第一个比较完整的数据库是 PRESAGE 数据库，它详细记录了每个蛋白质的实验状态、结构预测和建议[39]。还有一些结构基因组学联盟建立了在线的数据库，其中详细记录了他们目标蛋白的实验状态。比如说 ZebaView[40]、SPINE[41]和 ReportDB 数据库。对于结构基因组学而言，最大最全面的数据库是 TargetDB 数据库。它是 2001 年在 PRESAGE 数据库的基础上建立的。该数据库整合了来源于美国、加拿大、德国、以色列、日本、法国和英国的 28 个机构基因组中心的所有蛋白质数据。另外一个全面的数据库为蛋白表达纯化和结晶数据库(PepcDB)，它创建于 2004 年，是 TargetDB 数据库的延伸。该数据库收集了蛋白质结构解析过程中详细的实验状态和每一步的实验细节。还记录了实验过程，实验终止条件，可以重复利用的教程和美国 15 个结构基因组中心的交互信息[21]。随着结构基因组学的快速发展，有价值的实验数据积累的越来越多，为进行数据挖掘工作的研究者提供了充足的信息来提高预测方法的准确性。

3 构建预测模型

3.1 蛋白质可结晶性预测方法的输入特征

一般而言，蛋白质的特征及功能信息蕴藏在序列信息之中。所以我们可以用蛋白质的序列信息来作为预测方法的首选输入特征，用以计算其特性并预测可结晶性。

首先，最常用的蛋白质特征是等电点与疏水性。等电点和疏水性一般会影响蛋白质结晶过程的组装和折叠及相互作用等，所以对蛋白质的结晶性至关重要。蛋白质等电点的计算可以用 EMBOSS 软件包提供的 Bioperl 语言模块[42]来完成。而蛋白质的疏水性可以用 GRAVY 值表示，其计算方法为：每一个蛋白的 GRAVY 的值用序列中所有氨基酸的 Kyte-Doolittle 疏水值[43]的总和除以序列的长度。这两个特征都可以直接用相应的生物学软件获得[44]，非常方便快捷。

其次，用得较多的特征为蛋白质序列中氨基酸的组合，可以用组成向量[45]来表示氨基酸的组成。其中有二肽的组成、三肽的组成等。其计算方法都一样：20 种氨基酸依次按字母排列，表示为 AA_1、AA_2、…、AA_{19} 和 AA_{20}。序列中 AA_i 出现的次数记做 n_i，那么组成向量则可用下面公式来表示：

$$(\frac{n_1}{k}, \frac{n_2}{k}, \ldots\ldots, \frac{n_{19}}{k}, \frac{n_{20}}{k})$$

其中 k 为蛋白质序列的长度。

在特征表示时，可以这样理解：例如，如果某个氨基酸在序列中出现了 4 次，那么公式中 n_i 对应的值就为 4；如果该氨基酸在序列中没有出现，那么公式中 n_i 对应的值就为 0，以此类

推。由于氨基酸短范围的相互作用也可以影响蛋白的折叠[46]。因此还可以考虑氨基酸对被 P 个其他氨基酸所隔开的情况。可以考虑 P=0、1、2、3、4 五种情况。当 P=0 时，氨基酸对为二肽。其余的氨基酸对可以理解为二肽有缺口。也可以考虑三肽的组合及三肽之间被一个其他氨基酸隔开的情况。这样下来，蛋白质的特征会非常多，不利于计算、建模，而且准确率也不是很高。一般情况下，可以用特征选择方法来减少特征的幅度。常用的方法为：基于相关性特征子集的选择方法 (CFSS)。CFSS 通过考虑每个特征和它们之间的冗余度单独的预测能力评估了每个特征子集的值。搜索子集的策略是最佳优先搜索 (Best-first-search) 方法。该方法探索特定子集的间距用贪婪爬山算法 (Hill-climbing)。用该方法过滤后，所剩的特征不多且非常具有代表性，而且也大大提高了预测的准确性。

除此之外，蛋白质的物理化学信息 (分子量、疏水性、亲水性、折射率、平均可及表面积、柔韧性、熔点、侧链体积、侧链的疏水性、α 螺旋与 β 折叠的归一化率、极性、热容、等电点) 都可以用来对蛋白质可结晶性进行预测。

3.2 蛋白质结晶预测研究可利用的工具

用计算机模拟方法进行蛋白质可结晶性预测时，需要选择合适的分类器。简单的说，就是运用该分类器中的算法，将我们输入的可以影响蛋白质结晶的典型特征分为两类 (可结晶或者不可结晶)。当输入代表目标蛋白的新特征时，就可以预测出该蛋白质将要分到哪一类去。目前应用与蛋白质可结晶性预测的分类器有支持向量机 (Support vector machine, SVM)、怀卡托智能分析环境 (Waikato environment for knowledge analysis,WEKA)、R 软件包和自适应增强 (Adaptive boosting, AdaBoost) 分类器。

3.2.1 支持向量机 (SVM)

目前应用最广泛的蛋白质可结晶性预测工具是支持向量机 (SVM)[47]。SVM 是目前较为流行的分类软件。首先需要将数据库中的数据划分成训练集和测试集。训练集的每个特征，转化成 SVM 识别的格式 (目标值和属性，有相关软件可以直接转换)。SVM 的目标是基于训练数据产出一个模型 (Model)，用来预测只给出属性的测试数据的目标值。SVM 是基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。应用 SVM 的一般步骤为：1) 将数据转换成 SVM 格式包的格式；2) 对数据进行简单的缩放处理 (Scaling)；3) 考虑高斯径向基核函数 (Radial basis function kernel, RBF)；4) 使用交叉验证 (Cross validation) 寻找最佳参数 C 和 Y；5) 使用最佳参数 C 和 Y 来训练整个训练集；6) 测试。一般也可以用程序自带的脚本快速执行以上所有步骤。

3.2.2 怀卡托智能分析环境 (WEKA)

其次较为流行的预测蛋白质可结晶性软件是 WEKA[48]。WEKA 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。用户还可以通过 WEKA 的接口文档实现自己的数据挖掘算法。在使用 WEKA 时，

还是首先把我们的目标数据转换成可识别的格式，即一个二维的表格。表格里的一个横行称作一个实际特征，竖行称作一个属性 (Attribute)，相当于统计学中的一个变量。这样一个表格或者叫数据集，在 WEKA 看来，呈现了属性之间的一种关系 (Relation)。应用 WEKA，分为以下几步：1) 对目标数据进行预处理 (Preprocess)；2) 对输入特征进行分类 (Classify) 和聚类 (Cluster)；3) 用关联 (Associate)功能发掘前面导入的数据的隐藏关系；4) 选择属性 (Select attributes)；5) 结果可视化 (Visualize)。

3.2.3 R 程序包

还可以通过 R 程序包实现蛋白质可结晶性预测的功能。R 是一个程式语言和统计计算与绘图的整合环境。其语法与 S 语言 (S-Plus) 非常像，提供了非常多的统计工具。包括线性与非线性模型、统计检定、时间序列分析、分类分析、群集分析等相关工具。它具有很多优点：免费、开放、占有率高、跨平台、弹性大和互动式等。此方法不是很常用。

3.2.4 自适应增强分类器 (AdaBoost)

AdaBoost 也是一个应用广泛的分类工具，也称为 RAB。其核心思想是针对同一个训练集训练不同的分类器 (弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器 (强分类器)。使用 Adaboost 分类器可以排除一些不必要的训练数据特征，并将重心放在关键的训练数据上面。其一般步骤为：1) 给定训练样本集；2) 初始化样本；3) 迭代：训练样本的概率分布，训练弱分类器；计算弱分类器的错误率；选取合适阈值，使得错误率最小；更新样本权重；最终得到的强分类器。

4 评估构建模型的性能

当我们构建好预测模型后，就需要对该模型的性能进行评估。在生物信息学领域一般用准确率、特异性、敏感性和 MCC 4 个指标来评估预测模型的性能。在检验之前，我们首先要获得阳性数据 (可结晶的蛋白) 和阴性数据 (不可结晶的蛋白)。下面则是这 4 个指标的定义，灵敏度 (Sensitivity, Sn)：对于真实的数据，能够预测成“真的”比例；特异性 (Specificity, Sp)：对于阴性的数据，能够预测成“假”的比例；准确性 (Accuracy, Ac)：对于整个数据集 (包括阳性和阴性数据)，预测总共的准确比例；马修斯相关系数 (Mathew correlation coefficient, MCC)：当阳性数据的数量与阴性数据的数量差别较大时，能够更为公平地反映预测能力。它们的计算公式为：

$$\text{Accuracy} \quad \frac{\text{TP}+\text{TN}}{TP+TN+FP+FN}\times 100$$

$$\text{Sensitivity} = \frac{TP}{TP+FP}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{MCC} = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FN)}}$$

其中，真阳性 (TP)：阳性数据中被预测为阳性的数据；假阳性 (FP)：阴性数据中被预测为阳性的数据；真阴性 (TN)：阴性数据中被预测为阴性的数据；假阴性 (FN)：阳性数据中被预测为阴性的数据。

其中 MCC 的值越高，预测的可靠性越强。此外，还定义了接收者操作特征曲线 (Receiver-operator characteristics, ROC)。ROC 是对 TP rate

= $\frac{TP}{TP+FN}$ 和 FP rate $\frac{FP}{TN+FP}$ 作图得到的曲线，用 ROC 曲线下的面积 AROC 来评估预测的可靠性。AROC 越大，预测方法越可靠。

5 未来蛋白质可结晶性预测的展望

解析蛋白质结构具有重要的意义。但是实验方法进行蛋白质结晶研究工作费时费力，且不一定能得到满意的实验结果。计算机模拟方法可以提前对目标蛋白质进行相关分析，获得一定有用的结果 (等电点、疏水性、能否结晶和结晶沉淀剂等)。再加上网络数据库的不断完善，特别是 PSI 组织建立的 TargetDB 数据库。该数据库不但包含了成功的实验数据，也包含了失败的实验数据。失败的实验数据对于数据挖掘工作是非常重要的。若是没有这些数据，往往预测模型的正负数据不平衡问题就不好处理。因此，我们可以对这些数据库进行数据挖掘，找到一定的关系来构建预测模型，来预测有关蛋白质的结晶问题。目前已经利用这些资源发展起来很多蛋白质结晶预测的网络服务器。这些方法简单方便，能节省大量的时间和资源。并且预测也具有一定的可靠性，已经被广大研究者所接受，成功应用于结构生物学领域。随着计算机算法的发展、生物学软件的完善和蛋白质数据库的增长，计算机模拟的方法会越来越多，也越来越可靠。计算机模拟方法将会成为结构生物学领域不可缺少的一部分。

REFERENCES

[1] Bethel CM, Lieberman RL. Protein structure and function: an interdisciplinary multimedia-based guided-inquiry education module for the high school science classroom. J Chem Educ, 2014, 91(1): 52–55.

[2] Xue YZ, Li XX, Pang SL, et al. Efficacy and safety of computer-assisted stereotactic transplantation of human retinal pigment epithelium cells in the treatment of parkinson disease. J Comput Assist Tomo, 2013, 37(3): 333–337.

[3] Chen CY. A novel integrated framework and improved methodology of computer-aided drug design. Curr Top Med Chem, 2013, 13(9): 965–988.

[4] Cordeiro MN, Speck-Planche A. Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. Curr Top Med Chem, 2012, 12(24): 2703–2704.

[5] Gulerez IE, Gehring K. X-ray crystallography and NMR as tools for the study of protein tyrosine phosphatases. Methods, 2014, 65(2): 175–183.

[6] Tyszka JM, Fraser SE, Jacobs RE. Magnetic resonance microscopy: recent advances and applications. Curr Opin Biotech, 2005, 16(1): 93–99.

[7] Yee A, Pardee K, Christendat D, et al. Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. Accounts Chem Res, 2003, 36(3): 183–189.

[8] Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. Biochem Bioph Res Co, 2007, 355(3): 764–769.

[9] Sanchez-Puig N, Sauter C, Lorber B, et al. Predicting protein crystallizability and nucleation. Protein Peptide Lett, 2012, 19(7): 725–731.

[10] Smialowski P, Schmidt T, Cox J, et al. Will my protein crystallize? A sequence-based predictor. Proteins, 2006, 62(2): 343–355.

[11] Bourne PE, Westbrook JD, Berman HM, et al. The protein data bank (PDB) as a research tool. Abstr Pap Am Chem S, 2003, 226: U302–U302.

[12] Kuan TW, Wang JF, Wang JC, et al. VLSI design of an SVM learning core on sequential minimal optimization algorithm. Ieee T Vlsi Syst, 2012, 20(4): 673–683.

[13] Hernandez-Gonzalez J, Inza I, Lozano JA. Learning Bayesian network classifiers from label proportions. Pattern Recogn, 2013, 46(12): 3425–3440.

[14] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23(19): 2507–2517.

[15] Overton IM, Barton GJ. A normalised scale for structural genomics target ranking: the OB-Score. Febs Lett, 2006, 580(16): 4005–4009.

[16] Schmidberger M, Mansmann U. Parallel Computing with the R Language in a Supercomputing Environment. Berlin Heidelberg: Springer, 2009: 769–780.

[17] Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. Nucleic Acids Res, 2004, 32: D138–D141.

[18] Slabinski L, Jaroszewski L, Rychlewski L, et al. XtalPred: a web server for prediction of protein crystallizability. Bioinformatics, 2007, 23(24): 3403–3405.

[19] Overton IM, Padovani G, Girolami MA, et al. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. Bioinformatics, 2008, 24(7): 901–907.

[20] Chen L, Oughtred R, Berman HM, et al. TargetDB: a target registration database for structural genomics projects. Bioinformatics, 2004, 20(16): 2860–2862.

[21] Kouranov A, Xie L, De la Cruz J, et al. The RCSB PDB information portal for structural genomics. Nucleic Acids Res, 2006, 34: D302–D305.

[22] Kurgan L, Razib AA, Aghakhani S, et al. CRYSTALP2: sequence-based protein crystallization propensity prediction. Bmc Struct Biol, 2009, 9(1): 50.

[23] Kandaswamy KK, Pugalenthi G, Suganthan PN, et al. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. Protein Peptide Lett, 2010, 17(4): 423–430.

[24] Frank E, Hall M, Trigg L, et al. Data mining in bioinformatics using Weka. Bioinformatics, 2004, 20(15): 2479–2481.

[25] Hsieh CW, Hsu HH, Pai TW. Protein crystallization prediction with AdaBoost. Int J Data Min Bioin, 2013, 7(2): 214–227.

[26] Liu N, Wang H. Improving predictive accuracy by evolving feature selection for face recognition. Ieice Electron Expr, 2008, 5(24): 1061–1066.

[27] Niu B, Cai YD, Lu WC, et al. Predicting protein structural class with AdaBoost learner. Protein Peptide Lett, 2006, 13(5): 489–492.

[28] Gao JZ, Hu G, Wu ZH, et al. Improved prediction of protein crystallization, purification and production propensity using hybrid sequence representation. Curr Bioinform, 2014, 9(1): 57–64.

[29] Jahandideh S, Mahdavi A. RFCRYS: Sequence-based protein crystallization propensity prediction by means of random forest. J Theor Biol, 2012, 306: 115–119.

[30] Mizianty MJ, Kurgan LA. CRYSpred: accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics. Protein Peptide Lett, 2012, 19(1): 40–49.

[31] Mizianty MJ, Kurgan L. Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics, 2011, 27(13): I24–I33.

[32] Zucker FH, Stewart C, dela Rosa J, et al. Prediction of protein crystallization outcome using a hybrid method. J Struct Biol, 2010, 171(1): 64–73.

[33] Bellini D, Caly DL, McCarthy Y, et al. Crystal structure of an HD-GYP domain cyclic-di-GMP phosphodiesterase reveals an enzyme with a novel trinuclear catalytic iron centre. Mol Microbiol, 2014, 91(1): 26–38.

[34] Gomez-Garcia I, Stuiver M, Ereno J, et al. Purification, crystallization and preliminary crystallographic analysis of the CBS-domain pair of cyclin M2 (CNNM2). Acta Crystallogr F, 2012, 68: 1198–1203.

[35] Kumar A, Mocklinghoff S, Yumoto F, et al.

Structure of a novel winged-helix like domain from human NFRKB protein. PLoS ONE, 2012, 7(9): e43761.

[36] Lu HM, Yin DC, Liu YM, et al. Correlation between protein sequence similarity and crystallization reagents in the biological macromolecule crystallization database. Int J Mol Sci, 2012, 13(8): 9514–9526.

[37] Stevens RC. High-throughput protein crystallization. Curr Opin Struc Biol, 2000, 10(5): 558–563.

[38] Rodrigues A, Hubbard RE. Making decisions for structural genomics. Brief Bioinform, 2003, 4(2): 150–167.

[39] Brenner SE, Barken D, Levitt M. The PRESAGE database for structural genomics. Nucleic Acids Res, 1999, 27(1): 251–253.

[40] Wunderlich Z. ZebaView: a database tool for structural genomics. Undergraduate Laboratory Res, 2002, 694: 382–387.

[41] Bertone P, Kluger Y, Lan N, et al. SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res, 2001, 29(13): 2884–2898.

[42] Olson SA. EMBOSS opens up sequence analysis. European molecular biology open software suite. Brief Bioinform, 2002, 3(1): 87–91.

[43] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol, 1982, 157(1): 105–132.

[44] Chen K, Kurgan L, Ruan J. Optimization of the sliding window size for protein structure prediction. 2006 International Conference on Computational Intelligence in Bioinformatics and Computational Biology, 2006: 366–372.

[45] Chen C, Tian YX, Zou XY, et al. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol, 2006, 243(3): 444–448.

[46] Chen K, Kurgan L, Ruan J. Optimization of the sliding window size for protein structure prediction. IEEE, 2006: 1–7.

[47] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM TIST, 2011, 2(3): 27.

[48] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10–18.

(本文责编 陈宏宇)