DOI:10.12113/j.issn.1672-5565.201808002

蛋白质二级结构在线服务器预测评估

朱树平,刘毅慧*
(齐鲁工业大学(山东省科学院)计算机科学与技术学院,济南 250353)

摘 要:蛋白质二级结构的预测,对于研究蛋白质的功能和人类生命科学意义非凡。1951年开始提出预测蛋白质二级结构,1983年对于二级结构的预测只有50%的准确率。经过多年的发展,预测方式不断的改进和完善,到如今准确率已经超过80%。但目前预测在线服务器繁多,连续自动模型评估(CAMEO)也只给出服务器三级结构的预测评估,二级结构评估还未实现。针对上述问题,选取了以下6个服务器:PSRSM、MUFOLD、SPIDER、RAPTORX、JPRED和PSIPRED,对其预测的二级结构进行评估。并且为保证测试集不在训练集内,实验数据选取蛋白质结构数据库(Protein Data Bank,PDB)最新发布的蛋白质。在基于蛋白质同源性30%、50%和70%的实验中,PSRSM取得*Q*3的准确率分别为91.44%、88.12%和90.17%,比其他预测服务器中最高的MUFOLD分别高出3.19%、1.33%和2.19%,证明在同一类同源性数据中PSRSM比其他服务器有更好的预测效果。除此之外实验也得到其预测的*Sov*准确度也比其他服务器要高。比较各类服务器的方法与结果,得出今后蛋白质二级结构预测应当重点从大数据、模板和深度学习的角度进行研究。

关键词:蛋白质二级结构;预测;在线服务器;准确率;评估
中图分类号:Q518.1 **文献标志码**:A **文章编号**:1672-5565(2019)01-053-08

Protein secondary structure online server predictive evaluation

ZHU Shuping, LIU Yihui*
(*School of Computer Science and Technology*, *Qilu University of Technology* (*Shandong Academy of Sciences*), *Jinan* 250353, *China*)

Abstract: The prediction of protein secondary structure is of great significance for studying the function of proteins and human life sciences. The prediction of protein secondary structure was put forward in 1951, but the accuracy rate was only 50% in 1983. During years of development, the prediction method has been continuously optimized, and the accuracy rate has already exceeded 80%. However, there are many online servers, and Continuous Automate Model EvaluatiOn (CAMEO) can only provide predictive evaluation of the server's three-level structure, while the secondary structure evaluation has not been realized. Aiming to solve the above problems, PSRSM, MUFOLD, SPIDER, RAPTORX, JPRED, and PSIPRED were selected to evaluate their predicted secondary structure. The latest released protein from the Protein Data Bank (PDB) was applied to ensure that the test set is not included in the training set. In the experiments where the protein homology was 30%, 50% and 70%, the obtained accuracy of PSRSM for Q3 were 91.44%, 88.12%, and 90.17%, respectively. The accuracy was higher than the best prediction server MUFOLD by 3.19%, 1.33%, and 2.19% correspondingly, which proved that PSRSM has better prediction accuracy than other servers for the same kind of homology data and for the *Sov*. This paper focuses on analyzing the operating methods and corresponding results of various servers, thus it is concluded that the prediction of protein secondary structure should be studied from the perspectives of big data, templates, and in-depth learning.
Keywords: Protein secondary structure; Prediction; Online server; Accuracy; Evaluation

蛋白质是人体的有机大分子,是生命活动的主要承担者,在生物信息学领域,一直致力于对于蛋白

收稿日期:2018-07-13;**修回日期**:2018-12-21.
基金项目:国家自然科学基金(No.61375013);山东省自然科学基金(No.ZR2013FM020).
作者作简:朱树平,女,硕士研究生,研究方向:生物医学信息处理、智能计算.E-mail:Elvira29_zsp@163.com.
*通信作者:刘毅慧,女,教授,研究方向:生物计算、智能信息处理.E-mail:yxl@qlu.edu.cn.

质的研究。为了研究蛋白质的功能,往往从结构入手,但蛋白质结构有多种,其中关于二级结构的研究,有助于发现三维立体结构和提供蛋白质功能注解,因此大多数人都致力于蛋白质二级结构的研究。

在 1951 年,鲍林和科里首次提出了关于蛋白质二级结构问题[1],最初对于蛋白质二级结构的预测方法主要是通过研究氨基酸序列来进行,准确率在 60%左右。Rost[2-3] 等人在研究中采用 PHD 算法,把多序列排列中包含的进化信息作为神经网络的输入,预测蛋白质的二级结构准确率超过了 70%。Zafer[4] 等人使用动态贝叶斯分类器的稀疏算法,得到了 76.3%的准确率。Kurniawan[5] 等人使用 SVM 结合位置特异性打分矩阵(Position-specific scoring matrices,PSSM)和蛋白质结构的物理化学特征来预测,准确率达到 80%左右。Wang[6] 等人通过结合 PSSM 和氨基酸序列信息,并使用一种称为二级结构递归编码器-解码器网络(SSREDN)来解决输入蛋白质特征与 SS 之间的序列-结构映射关系,使用 CullPDB 和 CB513 数据库测试,分别达到 84.2%,82.9%的 *Q*3 准确率。蛋白质二级结构预测方式不断注入新的活力,现在很多方法都实现了在线服务器的预测,本文选取了 PSRSM、MUFOLD、SPIDER、RAPTORX、JPRED 和 PSIPRED 6 种服务器,分别阐述其算法原理,并通过测试数据比较每一个的预测准确度,从而给出当前在线服务器二级结构的评估。

1 在线服务器原理

1.1 PSRSM

该服务器使用基于数据分区和半随机子空间(Partition and semi-random subspace method,PSRSM)的方法[7]。在传统的随机子空间方法中,低维子空间是由高维空间随机采样产生的,PSRSM 使用的半随机子空间方法能够有效的保证基础分类器的准确性和多样化。该方法的主要步骤如下:首先把训练数据根据蛋白质的长度划分为不同的子集合,建立模型;然后使用半随机子空间的方法生成子空间,并在子空间上训练基础分类器;最后根据多数投票的规则,在子集上把分类器结合起来,生成最终的分类器,其中使用 SVM 作为最基本的分类器。

具体来说,对于输入使用 PSI-BLAST 程序生成 PSSM 数据,并且 PSI-BLAST 使用 BLOSUM62 进化矩阵搜索 NCBI 的非冗余(NR)数据库的缩减版本,按照上述原则得到的 PSSM 是 20 * L 的矩阵,20 为氨基酸的个数,L 为每个蛋白质的长度。在实验中使用 13 个滑动窗口来获取蛋白质序列信息和预测序列中心的蛋白质二级结构。假设输入一个长度为 L 的蛋白质,会产生 260 * L(13 * 20 * L)的输入矩阵。从 260 个特征值选取 160 个作为主要特征,作为网络输入。最后建立 12 个分类器进行训练。那么一个新的蛋白质序列会根据其长度,选择合适的分类器进行预测。

实验的训练集选取了 ASTRAL 数据集的 6 892 条蛋白质数据和 CullPDB 数据集的 12 288 条蛋白质数据,去掉相似度较高的蛋白质后,训练集总共包括 15 696 条数据。测试集使用 99 个 CASP10 数据、81 个 CASP11 数据、19 个 CASP12 数据、513 个 CB513 数据、1 673 个 25PDB 的数据和 2018 年 2 月 1 号之前的 100 条数据(T100),实验得到使用 6 个 GTPCs 模型在 25PDB、CB513、CASP10、CASP11、CASP12 和 T100 数据中的蛋白质二级结构的 *Q*3 预测准确率分别是 86.38%、84.53%、85.51%、85.89%、85.55%和 85.09%。该服务器预测蛋白质序列范围是 10 到 800,预测网址为:http://qilubio.qlu.edu.cn:82/protein_PSRSM/default.aspx。

1.2 MUFOLD

MUFOLD 采用的是一种名为深度初始-内部-初始(Deep 3I)的新型网络来预测蛋白质二级结构,并且对于输入的特征矩阵做了细致考量,特征矩阵中结合了氨基酸的理化性质、PSI-Blast 特征和 HHBlits 特征[8]。其中对于理化性质的特征矩阵,设置了从-1 到 1 之间选取的 8 个数字来表示一个氨基酸,前 7 位表示氨基酸理化性质,后一位用 1 或 0 表示是否输入氨基酸。如表 1,“*”表示某一类氨基酸,“n”表示依据理化性质设置的数值。MUFOLD 设置默认输入矩阵为 700 * 8,若假设输入一个氨基酸序列个数为 600 的蛋白质,设置矩阵时会把前 600 行的前 7 位按照本身理化性质设置,第 8 位设为 0,而后 100 行的前 7 为全部设为 0,后一位设置为 1。

对于 PSI-Blast 的特征,按照类似原理用从 0 到 1 的选取 21 位数字表示一个氨基酸,前 20 位根据得到的 PSSM 值设置,后一位用 1 或 0 表示是否有输入;对于 HHBlits 特征则用 0 到 1 之间的 31 位数字表示一个氨基酸,前 30 位根据 HMM 文件设置,最后一位同样用 0 或 1 表示输入。以上三个特征被组合成一个 58 位的特征,作为网络的输入。

Deep3I 网络是由 2 个 Deep3I 块、一系列卷积和完全联通的致密层构成。而 Deep3I 块是由初始模块递归嵌套构成,初始模块通过卷积操作能够有效提取氨基酸残基之间的非局部相互作用。Deep3I 网络通过用 TensorFlow 和 Keras 不断进行训练和实验来对蛋白质二级结构进行预测。

表 1　按照氨基酸理化性质设置的输入矩阵

Table 1　Input matrix set according to the physical and chemical properties of amino acids

输入个数	氨基酸类	7 位理化性质+1 位输入确认							
1	*	n	n	n	n	n	n	n	0
2	*	n	n	n	n	n	n	n	0
…	…	…	…	…	…	…	…	…	…
599	*	n	n	n	n	n	n	n	0
600	*	n	n	n	n	n	n	n	0
601	\	0	0	0	0	0	0	0	1
602	\	0	0	0	0	0	0	0	1
…	…	…	…	…	…	…	…	…	…
699	\	0	0	0	0	0	0	0	1
700	\	0	0	0	0	0	0	0	1

MUFOLD 实验中的数据集使用蛋白质序列长度介于 50 到 700 之间的数据,来自 CullPDB、JPRED、CASP、CB513 和 PDB 5 个公开的蛋白质数据库。具体来说:从 CullPDB 选取了9 581条数据,其中随机选出9 000条作为训练集,剩下的 581 条作为测试;从 JPRED 选取的数据均来自不同的超级家族;CASP 的数据集经过筛选后 CASP10 的 98 条数据,CASP11 的 83 条数据,CASP12 的 40 条数据被使用;CB513 和 385 条 PDB 数据也同样被应用于 MUFOLD 的实验中。MUFOLD 测试数据的范围是 30 到 700,测试网址是:http://mufold.org/mufold-ss-angle/。

1.3　SPIDER

Hefferman[9] 等人提到对于蛋白质二级结构预测和溶剂接触表面积的研究,多年一直停滞不前的原因来自于,有些氨基酸残基在三维结构中距离很近而在蛋白质序列中距离很远,因此较难捕获氨基酸残基之间的非局部相互作用。现有的机器学习的方法基本都使用 10~20 个滑动窗口来获取氨基酸的相互作用。而 SPIDER 不使用滑动窗口,采用一种长期短期记忆(Long Short-Term Memory , LSTM)双向递归神经网络(Bidirectional Recurrent Neural Networks ,BRNNs)的机器学习模型来实现预测,能够捕捉氨基酸残基之间的非局部相互相互作用,实验证明它能够改善蛋白质二级结构、骨干角度、接触号码和溶剂可及性的预测。

该网络的 LSTM-BRNN 模型是由两个使用 LSTM 细胞的 BRNN 层和两个紧密连接用整流线性单元(Rectified Linear Unit, ReLU)激活的隐含层构成,它被用于四次迭代中。对于该网络的输入,包含了 7 种具有代表性的蛋白质氨基酸理化性质(Physio-chemical properties, PP)、20 维来自 PSI-Blast 的 PSSM 和 30 维来自 HHBlits 每个残基的隐藏马尔科夫模型的序列谱(HMM Profiles),把这些数据放入由 LSTM-BRNNs 网络构成的迭代中,进行四次迭代(其中一次迭代包括两个 LSTM-BRNN),最后得到最终机器学习模型。该过程主要结构如图 1 所示。在训练期间为防止过拟合,使用丢失率为 50%的丢失算法,并用 Adam 优化训练过程,该网络能够在不使用滑动窗口的条件下捕获长短距离交互。

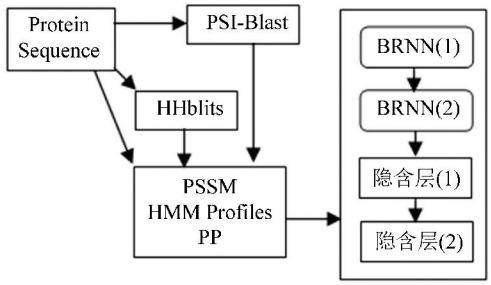

图 1　SPIDER 主要结构

Fig.1　Main structure of SPIDER

1.4　RAPTORX

RAPTORX 使用由深度卷积神经网络(Deep convolutional neural network , DCNN)和条件随机场(Conditional random fields,CRF)组合而成的深度卷积神经场(Deep Convolutional Neural Fields,DCNF),来预测蛋白质二级结构,并且对网络采用一种在 ROC 曲线下面积的(Area under the ROC curve, AUC)最大化方法来训练,从而能够很好地解决紊乱序列蛋白质的预测问题[11]。Wang[12] 提到在使用蛋白质序列文件后,RAPTORX 在数据集 CASP 和 CAMEO 能够得到大约为 84%的 $Q3$ 准确率和 72%的 $Q8$ 准确率,不使用序列文件能够获得约为 74%

的 *Q*3 准确率和 59%的 *Q*8 准确率,它能够有效的解决复杂的基因结构关系建模和相邻残基间的建模。Wang[13]指出 DCNF 使用 DCNN 代替 CNF 中使用的浅层神经网络,能够捕获输入和输出标签之间复杂的关系,并且能够捕获远程的序列信息。

RAPTORX 实验中使用的数据有 6 125 个 CullPDB 数据,CB513 数据、123 个 CASP10 数据、105 个 CASP11 数据和 CAMEO 的数据,还有 JPRED 公开的 1 338 个训练数据和 149 个测试数据。RAPTORX 测试数据范围是 26 到 4 000 个蛋白质序列,预测网址为:http://raptorx.uchicago.edu/StructurePropertyPred/predict/。

1.5 JPRED

JPRED 服务器从 1998 年开始提供蛋白质的预测到现在已经发展到 JPRED4 版本。JPRED3 版本用 JNET 算法提供单个蛋白质序列或者多序列比对(MSA)的预测,其中 JNET 使用 JNET v2.0。JNET v2.0 不使用频率文件,只使用 PSI-BLAST 的 PSSM 配置文件和 HMMER 的隐马尔可夫模型,把神经网络由 9 个单元增加到 100 个单元,该方法是通过对超家族级别的 SCOPe 数据的 Astral 汇编衍生的序列和结构非冗余数据集进行 7 倍交叉验证培训而开发的[14],最后使用 149 条盲数据进行测试得到了 81.5%的 *Q*3 准确率。

JPRED4 版本和 JPRED3 一样,同样使用 JNET 算法并提供单一序列和多序列比对的蛋白质序列的二级预测。不同的是它选取 1 358 个 SCOPe/ASTRAL v.2.04 超级家族中的一个为代表,用 JNET 2.3.1 进行 7 倍交叉验证的实验,通过寻找 UniRef90 v.2014_07 来生成 PSI-BLAST 文件并为每一个蛋白质序列建立多重序列比对。最后在 150 个训练集上获得了 82%的准确率[15]。同时 JPRED 在线服务器也可以提供溶剂可及性和卷曲螺旋区的预测,预测网址为:http://www.compbio.dundee.ac.uk/jpred4/index.html。

1.6 PSIPRED

Mcguffin[16]等人指出 PSIPRED 服务器结合了三种先进的技术,分别是 PSIPRED、GenTHREADER 和 MEMSAT 2。其中 PSIPRED 采用严格的交叉验证程评估性能,并且采用两个前馈的神经网络,对从 PSI-BLAST 获得的输出进行分析,从而得到可靠的二级结构预测结果;GenTHREADER 用来推断跨膜蛋白的结构和拓扑结构;MEMSAT2 能够快速识别蛋白质的折叠信息,预测网址为:http://bioinf.cs.ucl.ac.uk/psipred/。

从以上 6 个服务器预测过程的角度分析,可以看到每个服务器各有优缺点。其中能够批量上传和下载实验结果的是 PSRSM、SPIDER3 和 RAPTORX,给定结果为压缩包的形式,需进一步整合。服务器 JPRED 和 PSIPRED 都必须遵循每次只能上传一个蛋白质文件(或序列)的约定,而且结果是以邮件的形式发送到邮箱里面,并且 PSIPRED 在同一时间段内最多只允许上传 20 条数据进行预测,因此预测结果获取过程较为复杂。MUFOLD 虽然网站上说明一次可以批量上传少于 10 条的数据但是在实验中获取数据,最多一次只可上传 4 条数据进行预测。6 个服务器预测的时间相差并不是很大,主要在于预测结果的获取方式上存在很大差距。

2 数据选取和评估标准

基于每个服务器都可以预测为前提,依据蛋白质发布的月份和其同源性分别选取了 150 条数据进行实验,并采用了合适的评价标准来评估。

2.1 数据选取

数据选取遵循以下原则:数据选取 2018 年 PDB 最新发布的数据,保证了测试集不在服务器的训练集中;数据来自不同的时间段,更具有分散性;数据量较大,使得实验结果更具有说服力;选取的蛋白质长度能够让每一个服务器都可以进行测试,并得到预测结果。基于上述的条件从 2018 年 4、5、6 月份分别选取了 50 条蛋白质序列进行第一次实验,数据选取如表 2 所示。

并且为了使实验结果更具有可靠性,又进一步从 2018 年 4 到 8 月,基于同源性的 30%,50%和 70%随机分别选取了 50 条数据,共 150 条数据(T150)进行第二次实验,该实验的数据选取如表 3 所示。

2.2 评估标准

本文采用了两种衡量蛋白质二级结构预测准确性方法:*Q*3 和 *Sov* 的值主要是衡量个别残基分配的精度,*Sov* 的值主要是衡量全元素的预测精度。

2.2.1 *Q*3

按照 DSSP[17]的规定,通常我们把蛋白质二级结构划分为 H、G、I、E、B、T、S 和-,8 种状态。而这 8 这种状态,按照 H、G、I→H,E、B→E,其他→C 的方式,将一条氨基酸序列转化为 H(螺旋)、E(折叠)、C(卷曲),3 种状态。则 *Q*3 表示被正确预测的三种状态的氨基酸数占整个氨基酸序列的比例。符合以下计算公式:

$$Q3 = \frac{S_H + S_E + S_C}{S} \tag{1}$$

其中: S_E 是 E 类蛋白质结构准确预测的数量, S_H 是 H 类蛋白质结构准确预测的数量, S_C 是 C 类蛋白质结构准确预测的数量, S 是指总的氨基酸数量, $Q3$ 指的是三种状态下,蛋白质二级结构预测的准确率。

表 2　DB150 数据集

Table 2　DB150 data set

发布时间/月	蛋白质名称									
4	5MXB	5MXW	5NFX	5NM3	5NWN	5NYH	5NZG	5NZH	5NZI	5NZJ
	5NZK	5NZI	5NZM	5O6C	5OAO	5ODX	5OHU	5OI7	5OI9	5OID
	5OLN	5OQH	5TOS	5USB	5USN	5USO	5V80	5VAS	5VCK	5VFX
	5VFY	5VFZ	5VG5	5VGC	5VGP	5VH2	5VHU	5VJ3	5VKX	5VYY
	5WBS	5WC7	5WDJ	5NQ0	5O8M	5OAE	5OET	5OLM	5X9B	5YQ5
5	5LTL	5MHA	5O1A	5O1F	5OBU	5OBW	5OBX	5OBY	5OU0	5OUJ
	5OUK	5SVH	5TF7	5TF8	5TFA	5TGK	5TRZ	5TS1	5TXK	5UK7
	5UNP	5VP3	5WKR	5WKS	5WL3	5WL5	5WL6	5WL7	5WNI	5WNK
	5WNL	5XK2	5XKC	5YO3	5YRP	5ZEO	6AU6	6BWV	6BXD	6BXE
	6BY2	6BY3	6CAM	6CCG	6CDF	6CDR	6CEN	6CEV	5OFE	5VWG
6	5M0B	5O42	5O6X	5O72	5OND	5TPT	5VZN	5VZQ	5W0Y	5W1D
	5W37	5WAA	5WLS	5WMO	5WMQ	5X84	5XPL	5XPQ	5XPS	5XQ5
	5XQA	5XWS	5XZK	5Y0F	5YD5	5YGU	5YK9	5YWR	5Z43	5ZJ6
	5ZK0	6BG6	6BVP	6CD3	6CMN	6CO2	6CPL	6CPN	6D0L	6D0S
	6D2C	6D60	6D8J	6D9N	6D9Y	6DIP	6F51	6F63	6F66	5O6P

表 3　T150 数据集

Table 3　T150 data set

同源性比例%	蛋白质名称									
30	5LOS	5LTL	5M6Y	5MCT	5MCU	5MCV	5MCW	5MF7	5MGS	5MH5
	5MH6	5MIY	5MLP	5MNW	5MQX	5MV2	5MXB	5MXP	5MLQ	5N12
	5N2P	5N9B	5NAP	5NBC	5NCB	5NCM	5NDX	5NFX	5NKN	5NM3
	5NPN	5NQ0	5NUK	5NV9	5NWH	5NYH	5NZ4	5NZ5	5NZG	5O9X
	5OBA	5OC9	5OD3	5OGX	5OOX	5OQS	5OTU	5OW5	5OWO	5QIF
50	5MQX	5MXP	5NCB	5NDX	5NZ4	5O9X	5OC9	5OD3	5OGX	5OOX
	5OQS	5OTU	5OW5	5OWB	5OWL	5OWO	5QIF	5UMP	5UMW	5VO3
	5VX5	5W6K	5W6Q	5W7R	5W87	5W8Y	5W8Z	5WA5	5WCH	5WCQ
	5WCX	5WD6	5WDG	5WDK	5WEC	5WEW	5WFU	5WG8	5WH5	5WH6
	5WJM	5WKW	5WLP	5WM9	5WQM	5X16	5XEJ	5XGQ	5XKE	5XOR
70	5OG7	5OG9	5OGH	5OGJ	5OGO	5OGT	5OGZ	5OHT	5OHY	5OIO
	5OII	5OIU	5OJ3	5OJ5	5OK2	5OK3	5OK6	5OLT	5OMI	5ONK
	5ONN	5ONQ	5OO3	5OP3	5OW4	5OWC	5OWK	5Q22	5ULY	5UQ9
	5VJA	5VTL	5WAX	5WID	5WMG	5WMK	5WNF	5WOE	5WP1	5WYN
	5X2I	5XFJ	5XNC	5XNE	5XQZ	5XWX	5XX0	5XXJ	5XXQ	5XYB

2.2.2　*Sov*

Sov 的计算是基于重叠片段比值的一种测度,它对预测结果和观察到的结果同等对待。同样按照上述 $Q3$ 的思想把蛋白质二级结构划分为螺旋、折叠和卷曲三种状态。如果假设观察到的序列记为 S_1 ,预测到的序列记为 S_2 , S_0 为 S_1 和 S_2 所有状态相同的片段,那么 S_0 必定会包含一对重叠和一个螺旋,接下来 S_1 的长度为 length (S_1),并且把每对中 S_1 和 S_2 序列个数求并集记为 max (S_1 , S_2),把 S_1 和 S_2 的序列个数求交集记为 min (S_1 , S_2)。在上述基础

上把 *Sov* 的计算公式定义为[18]：

$$Sov = \frac{100}{N_{sov}} \sum_{S_0} \left[\frac{\min(S_1,S_2) + \delta(S_1 + S_2)}{\max(S_1,S_2)} \text{length}(S_1) \right] \quad (2)$$

其中关于 δ 的设定是为了允许蛋白质结构中边缘处片段的变化，$\delta(S_1,S_2)$ 取值符合以下定义：

$$\delta(S_1,S_2) = \min \begin{Bmatrix} (\max(S_1,S_2) - \min(S_1,S_2)) \\ \min(S_1,S_2) \\ \text{int}[\text{length}(S_1) \div 2] \\ \text{int}[\text{length}(S2) \div 2] \end{Bmatrix} \quad (3)$$

3 实验及结果

从 PDB 中下载得到最新的蛋白质数据，然后分别上传到 6 个预测服务器上进行测试。上传蛋白质序列得到的预测结果后，通过与正确的三态的 DSSP 结果相比较，计算每一条蛋白质的 *Q*3 和 *Sov* 准确率。第一次实验中每月数据和 DB150 的 *Q*3 和 *Sov* 准确率如表 4 所示。第二次实验中基于 30%，50%，70%的同源度数据和 T150 的 *Q*3 和 *Sov* 的实验结果如表 5 所示。

表 4 实验 1 的 Q3 和 Sov 平均准确率

Table 4 Average accuracy of Q3 and Sov in Experiment 1

评估参数	数据集	MUFOLD/%	SPIDER3/%	PSRSM/%	RAPTORX/%	JPRED/%	PSIPRED/%
*Q*3	4 月	85.67	84.94	87.39	82.21	78.31	78.99
	5 月	88.48	86.70	87.66	85.23	80.40	80.44
	6 月	87.46	86.30	89.15	83.97	79.91	80.61
	DB150	87.20	85.98	88.07	83.80	79.54	80.02
Sov	4 月	79.67	81.35	81.61	76.39	72.91	72.25
	5 月	83.52	83.38	84.00	80.19	75.76	73.27
	6 月	83.53	82.34	83.36	78.84	74.91	73.88
	DB150	82.24	82.35	82.99	78.47	74.53	73.13

表 5 实验 2 的 Q3 和 Sov 平均准确率

Table 5 Average accuracy of Q3 and Sov in Experiment 2

评估参数	数据集	MUFOLD/%	SPIDER3/%	PSRSM/%	RAPTORX/%	JPRED/%	PSIPRED%
*Q*3	30%	88.25	87.44	91.44	84.20	80.88	82.02
	50%	86.79	85.87	88.12	84.54	79.68	80.53
	70%	87.98	86.19	90.17	83.45	78.62	80.15
	T150	87.67	86.50	89.91	84.07	79.73	80.90
Sov	30%	84.56	83.76	87.45	81.23	78.39	77.32
	50%	81.67	80.23	81.95	78.74	75.30	76.05
	70%	78.15	78.33	83.36	75.37	67.43	68.81
	T150	81.46	80.77	84.25	78.45	73.71	74.06

从实验结果中看到，不论是基于月份的蛋白质数据，还是基于同源性不同划分的数据，PSRSM 都取得了在同一类别中较好的效果，*Q*3 的预测准确率有时甚至超过 90%。按照月份划分时，4 月份的数据集中，PSRSM 达到了最好的预测效果，*Q*3 和 *Sov* 的值分别为 87.39% 和 81.61%；在 5 月份数据集中，MUFOLD 的 *Q*3 准确率最高，为 88.48%，*Sov* 准确率仅次于 PSRSM 的 84.00%，为 83.52%；在 6 月份数据集中 PSRSM 的 *Q*3 获得最高准确率为 89.15%，而 *Sov* 仅次于 MUFOLD 的 83.53%，为 83.36%。在综合数据 DB150 的结果中我们得到 6 种预测方式 *Q*3 的准确率由高到低为 PSRSM 的 88.07%，MUFOLD 的 87.20%，SPIDER 的 85.98%，RAPTORX 的 83.80%，PSIPRED 的 80.02% 和 JPRED 的 79.54%；*Sov* 准确率由高到低为 PSRSM 的 82.99%，SPIDER3 的 82.35%，RAPTORX 的 78.47%，JPRED 的 74.53% 和 PSIPRED 的 73.13%，PSRSM 得到了 *Q*3 和 *Sov* 的最高准确率。

在基于同源性的实验中,结果显示基于 30%时,PSRSM 得到了 91.44%的 *Q*3 准确度和 87.45%的 *Sov* 准确度,比其他服务器中最好的 MUFOLD 分别高出 3.19 和 2.89 个百分分点;同源度为 50%时,PSRSM 的 *Q*3 为 88.12%, *Sov* 为 81.95%,分别比 MUFOLD 高出 1.33 和 0.28 个百分点;70%的同源度时 PSRSM 的 *Q*3 和 *Sov* 分别为 90.17%和 83.36, *Q*3 比其他服务器中最好的 MUFOLD 高出 2.19%,Sov 比预测结果最好的 SPIDER 高出 5%。总体来看在 T150 中 *Q*3 和 Sov 准确率由高到低分别为 PSRSM 的 89.91 和 84.25%,MUFOLD 的 87.67%和 81.46%,SPIDER 的 86.50%和 80.77%,Raptorx 的 84.07%和 78.45%,PSIPRED 的 80.06%和 74.06%,JPRED 的 79.73 和 73.71%。

无论在哪一种情况下,PSRSM、MUFOLD 和 SPIDER3 都得到了超过 84.9%的 *Q*3 准确率和超过 78.1%的 *Sov* 准确率,其中 PSRSM 表现出良好的预测性能。

4　结　论

蛋白质二级结构预测的准确度,将决定人类对于蛋白质功能的了解程度。本文介绍了现在 6 个热门的预测服务器原理,并使用最新的数据对其二级结构预测的准确率进行评估。比较 6 个服务器的预测方法和实验结果,可以看到它们的研究方法都在着重解决那些三维结构中距离近而序列中距离远的氨基酸残基的预测问题,并为此一再提出新的解决思路。

PSRSM 在上述实验数据中大多都取得了最好的实验结果,特别是在基于同源性差异的实验中,当同源度较很低为 30%时,其 *Q*3 准确率比其他服务器中最好的 MUFOLD 高出 3.19%,这更说明 PSRSM 具有更好的预测效果。PSRSM 与其他服务器比较,其优点在于基于蛋白质长度划分设计模板的使用,另一点在于训练数据量非常庞大,当然也采用了合理的预测方法。通过该实验和结果也可以看出,其他服务器能否获得优越的结果与其训练数据量的大小密切相关,当然还与其各自使用的深度学习算法有关。因此今后对于蛋白质二级结构预测的研究应当重点从大数据、模板和深度学习的角度进行突破。

参考文献(References)

[1]YANG Yuedong, GAO Jianzhao, WANG Jihua, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch[J]. Briefings in Bioinformatics, 2018, 19(3): 482-494. DOI: 10.1093/bib/bbw129.

[2]ROST B, SANDER C. Prediction of protein secondary structure at better than 70% accuracy[J]. Journal of Molecular Biology, 1993, 232(2): 584-99. DOI: 10.1006/jmbi.1993.1413.

[3]ROST B, SANDER C. Combining evolutionary information and neural networks to predict protein secondary structure[J]. Proteins Structure Function & Bioinformatics, 1994, 19(1): 55-72. DOI: 10.1002/prot.340190108.

[4]ZAFER A, AJIT S, JEFF B. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure[J]. BMC Bioinformatics, 2011, 12(1): 154. DOI: 10.1186/1471-2105-12-154.

[5]KURNIAWAN I, HARYANTO T, HASIBUAN L S, et al. Combining PSSM and physicochemical feature for protein structure prediction with support vector machine[J]. Journal of Physics Conference Series, 2017, 835(1): 012006. DOI: 10.1088/1742-6596/835/1/012006.

[6]WANG Yangxu, MAO Hua, ZHANG Yi. Protein secondary structure prediction by using deep learning method[J]. Knowledge-Based Systems, 2017, 118(2): 115-123. DOI: 10.1016/j.knosys.2016.11.015.

[7]MA Yuming, LIU Yihui, CHENG Jinyong. Protein secondary structure prediction based on data partition and semi-random subspace method[J]. Scientific Reports, 2018, (8): 9856. DOI: 10.1038/s41598-018-28084-8.

[8]FANG Chao, SHANG Yi, XU Dong. MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2018, 86(5): 592-598. DOI: 10.1002/prot.25487.

[9]HEFFERMAN R, YANG Y, PALIWAL K, et al. Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers, and solvent accessibility[J]. Bioinformatics, 2017, 33(18): 3842-3849. DOI: 10.1093/bioinformatics/btx218.

[10]HEFFERNAN R, DEHZANGI A, LYONS J, et al. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins[J]. Bioinformatics, 2015, 32(6): 843-849. DOI: 10.1093/bioinformatics/btv665.

[11]WANG S, MA J Z, XU J B. AUCpreD: Proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields[J]. Bioinformatics, 2016, 32(17): i672-i679. DOI: 10.1093/bioinformatics/btw446.

[12]WANG S, LI W, LIU S W, et al. RaptorX-Property: a web server for protein structure property prediction[J]. Nucleic Acids Research, 2016, 44(W1): W430-W435. DOI: 10.1093/nar/gkw306.

[13]WANG S, PENG J, MA J Z, et al. Protein secondary

structure prediction using deep convolutional neural fields [J]. Scientific Reports, 2016, (6): 18962. DOI: 10.1038/srep18962.

[14] COLE C, BARBER J D, BARTON G J. The Jpred 3 secondary structure prediction server [J]. Nucleic Acids Research, 2008, 36(W12): W197-W201. DOI: 10.1093/nar/gkn238.

[15] DROZDETSKIY A, COLE C, PROCTER J, et al. JPred4: a protein secondary structure prediction server [J]. Nucleic Acids Research, 2015, 43(1): W389-W394. DOI: 10.1093/nar/gkv332.

[16] MCGUFFIN L J, BRYSON K, JONES D T. The PSIPRED protein structure prediction server [J]. Bioinformatics, 2000, 16(4): 404-405. DOI: 10.1093/bioinformatics/16.4.404.

[17] WAKAMURA K, HIROKAWA K, ORITA K. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features [J]. Biopolymers, 1983, 22(12): 2577-2637. DOI: 10.1002/bip.360221211.

[18] 泽瓦勒贝 M, 等. 理解生物信息学 [M]. 李亦学, 郝沛, 译. 北京: 科学出版社, 2012.
ZERLEBIL M, et al. Understanding bioinformatics [M]. LI Yixue, HAO Pei, Trans. Beijing: Science Press, 2012.