DOI:10.12113/201907006

蛋白质二级结构预测服务器 PSRSM

韩心怡,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 250300)

摘 要:蛋白质二级结构预测是蛋白质结构研究的一个重要环节,大量的新预测方法被提出的同时,也不断有新的蛋白质二级结构预测服务器出现。试验选取7种目前常用的蛋白质二级结构预测服务器:PSRSM、SPOT-1D、MUFOLD、Spider3、RaptorX,Psipred和Jpred4,对它们进行了使用方法的介绍和预测效果的评估。随机选取了PDB在2018年8月至11月份发布的180条蛋白质作为测试集,评估角度为:Q3、Sov、边界识别率、内部识别率、转角C识别率,折叠E识别率和螺旋H识别率七种角度。上述服务器180条测试数据的Q3结果分别为:89.96%、88.18%、86.74%、85.77%、83.61%,79.72%和78.29%。结果表明PSRSM的预测结果最好。180条测试集中,以同源性30%,40%,70%分类的实验结果中,PSRSM的Q3结果分别为:89.49%、90.53%、89.87%,均优于其他服务器。实验结果表明,蛋白质二级结构预测可从结合多种深度学习方法以及使用大数据训练模型方向做进一步的研究。

关键词:蛋白质;蛋白质二级结构预测;PSRSM;预测方法评估
中图分类号:Q518.1 **文献标志码**:A **文章编号**:1672-5565(2020)02-116-11

Protein secondary structure prediction Server PSRSM

HAN Xinyi, LIU Yihui*
(*School of Computer Science and Technology*, *Qilu University of Technology* (*Shandong Academy of Sciences*), *Jinan* 250300, *China*)

Abstract: Protein secondary structure prediction is an important part of protein structure research. When a large number of new prediction methods are proposed, new protein secondary structure prediction servers emerge. This paper selects seven commonly used protein secondary structure prediction servers: PSRSM, SPOT-1D, MUFOLD, Spider3, RaptorX, Psipred, and Jpred4 to evaluate their instructions and predicted effects. The evaluation data set is the 180 proteins released by the randomly selected PDB from August to November in 2018. The evaluation parameters are Q3, Sov, boundary recognition rate, internal recognition rate, corner C recognition rate, folding E recognition rate, and spiral H recognition rate. The Q3 results of the above servers were 89.96%, 88.18%, 86.74%, 85.77%, 83.61%, 79.72%, and 78.29%, which show that the prediction results of PSRSM were the best. In the 180 test sets, the results of the classification of 30%, 40%, and 70% homology show that the Q3 results of PSRSM were 89.49%, 90.53%, and 89.87%, respectively, which were superior to other servers. The experimental results suggest that protein secondary structure prediction could be further studied by combining multiple deep learning methods and using the big data training model.

Keywords: Protein; Protein secondary structure prediction; PSRSM; Prediction method evaluation

蛋白质二级结构[1]预测是生物信息学领域一项非常重要的研究课题,蛋白质二级结构不仅是构成蛋白质稳定构象的基础,同时也是进一步研究蛋白质三级结构的重要环节[2]。随着越来越多的蛋白质序列顺利完成了结构测试,国际上也不断有新的蛋白质二级结构预测方法被提出,同时也提供了

收稿日期:2019-07-26; **修回日期**:2019-10-11.
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)资助.
作者简介:韩心怡,男,硕士研究生,研究方向:生物医学信息处理、智能计算. E-mial: hanxinyi1234@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.

多种在线预测服务器。试验选取了七种在线预测服务器:PSRSM、Spider3、SPOT-1D、RaptorX、MUFOLD, Psipred 和 Jpred4,并将它们的预测结果从 Q3、Sov、边界识别率、内部识别率、转角 C 识别率,折叠 E 识别率和螺旋 H 识别率七个方面进行了对比评估。上述七种在线预测服务器,均采用了各自不同的预测方法:PSRSM 采用基于数据划分和半随机子空间的预测方法[3];Spider3 使用长短时记忆网络和双向递归神经网络的混合模型[4];SPOT-1D 结合了残余卷积网络和双向递归神经网络[5];RaptorX 使用了深度卷积神经场[6];MUFOLD 采用了一种名为深度初始-内部-初始的网络[7];Jpred4 通过 JNet[8] 算法提供预测,还有使用前馈神经网络的 Psipred[9]。最新出现的 PSRSM 和 SPOT-1D 也增加了对大数据集的使用。

相比于文献[10],增加了对最新发布的 SPOT-1D 服务器介绍和评估,对所有服务器的使用流程做出了说明,同时增加了对转角 C、折叠 E 和螺旋 H、内部和边界结构的预测准确率评估,为研究者提供更多的参考角度。其中,各服务器 Q3 结果从高到低分别为 PSRSM:89.96%;SPOT-1:88.18%;MUFOLD:86.74%;SPIDER3:85.77%;RaptorX:83.61%;Psipred:79.72%;Jpred4:78.29%。结果表明 PSRSM 预测效果优于其他服务器。

1　PSRSM-Server

PSRSM-Server 是由齐鲁工业大学智能信息处理团队开发的蛋白质二级结构预测服务器,该服务器基于数据划分和半随机子空间(Partition and semi-random subspace, PSRSM)方法进行预测[3]。方法的主要流程为:首先根据蛋白质序列的长度将训练集划分为 6 种子集,然后用半随机子空间方法生成子空间,将 SVM 作为基本分类器,在子空间中训练基本分类器;最后通过多数投票规则把子集中的基本分类器结合,生成最终的分类器。网络输入为 PSI-BLAST 程序生成的 20×L 的 PSSM 矩阵,其中 20 为氨基酸个数,L 为蛋白质长度。输入的蛋白质序列将会根据长度选择合适的分类器进行预测。此服务器将预测结果根据“H、G、I 转为 H”,“B、E 转为 E”,“其他结构转为 C”的规则得出最终的 3 态结果。该方法在 ASTRAL 和 CullPDB 数据集上选取了 15 696条去除较高相似度的数据上进行训练,在测试集 CASP10、CASP11、CASP12、CB513, 25PDB 和 T100(2018 年 2 月前的 100 条)上 Q3 识别率分别达到 85.51%、85.89%、85.55%、84.53%, 86.38% 和 85.09% 的良好性能[3]。PSRSM-Server 网址为:http://210.44.144.20:82/protein_PSRSM/default.aspx。

该网站提供了单条序列预测和批量序列预测的功能,点击“Sequence”,按照图 1 所示,输入邮箱,便可进行单条作业提交。所支持的蛋白质长度范围为 10~800。

提交成功后网站会分配一个 Job ID,使用者可根据此 Job ID、序列或者预留邮箱在网站左侧“Predicted result”中根据不同的方式进行结果查询,如图 2 所示。

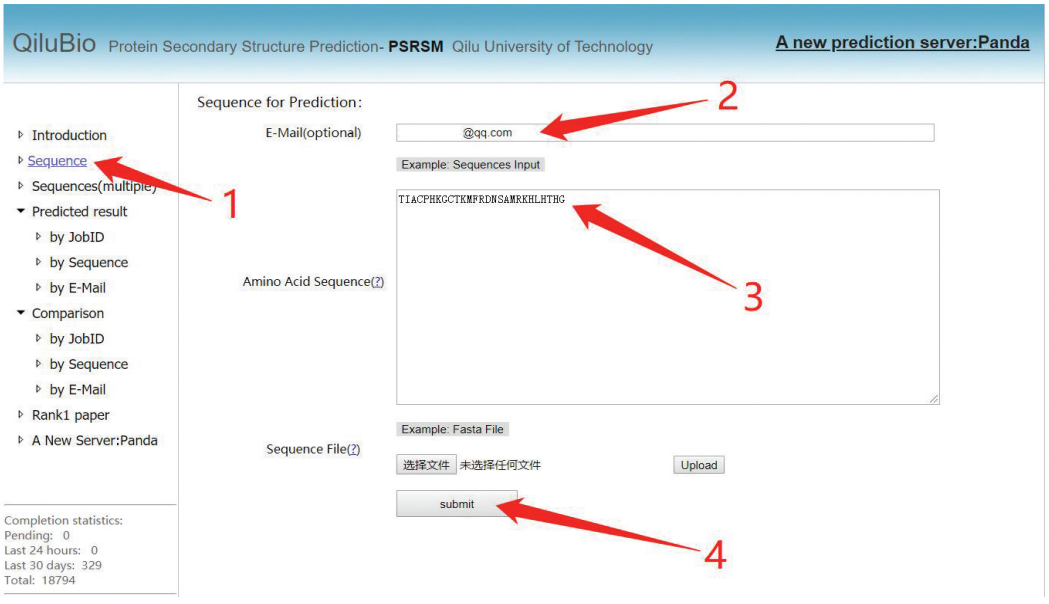

图 1　PSRSM 服务器单条数据测试提交步骤

Fig.1　Single data test submission step of PSRSM-Server

注:用户可根据图中标注 1 点击“Sequence”然后输入查询邮箱,在标注 3 处输入提交序列,最后点击标注 4 处的“submit”完成单条数据提交.

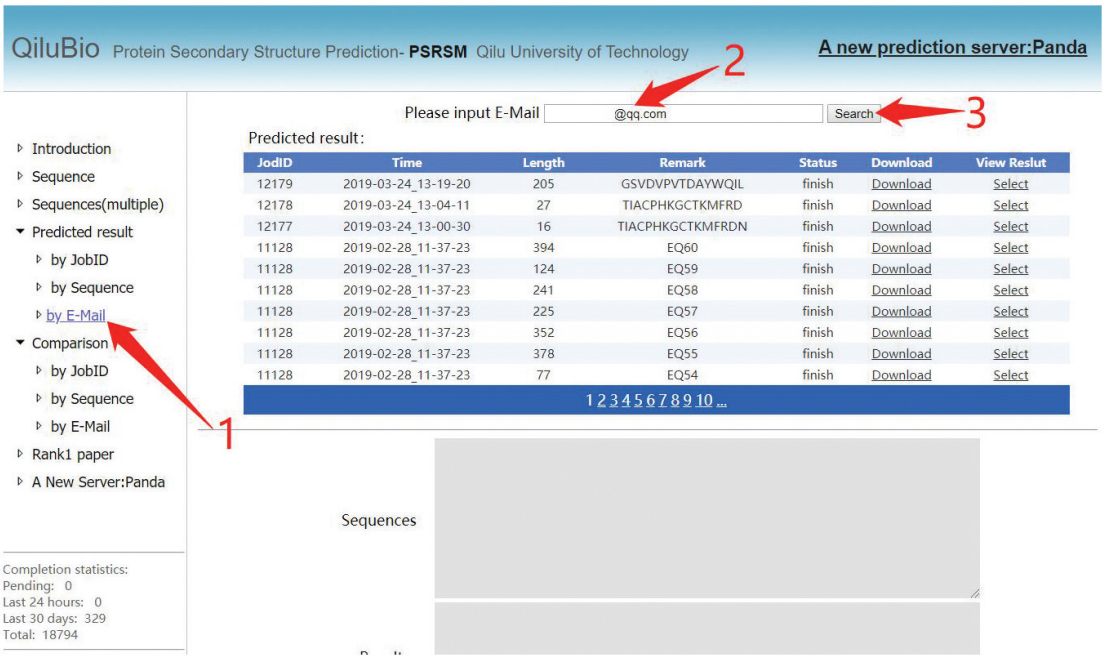

图 2 PSRSM 服务器根据预留邮箱结果查询

Fig.2 Query results based on email address in PSRSM-Server

注:用户首先在“Predicted result”处点击“by E-Mail”,然后在标注 2 处输入图 1.2 输入的预留邮箱,最后点击标注 3 处的 Search 即可查询测试状态.

用户可根据需要,选择点击“Download”下载结果或者点击“Select”在网页端查看结果,下载的结果将以 txt 格式保存。网站也同样支持上传 Fasta 格式文件进行预测,查询结果方式同图 2。最后,该网站提供了查询预测准确率的功能,在左侧“Comparison”中,选中所需查询的结果,输入真实的 DSSP,可直接查看 Q3 和 Sov 准确率。

2 其他方法介绍

2.1 Spider3

现有的机器学习方法在预测蛋白质二级结构时通常依赖于设置 10 到 20 个氨基酸残基大小的滑动窗口来捕捉“短到中”距离的残基相互作用,而该方法基于长短时记忆(Long short-term memory, LSTM)双向递归神经网络(Bidirectional recurrent neural network, BRNNs)[4],在不设置滑动窗口的情况下捕捉长距离的残基交互,改善了蛋白质二级结构的预测效果。该方法模型使用了两个节点数为 256 的双向递归神经网络层(BRNN),之后为两层节点分别为 1 024和 512 的隐藏层。在 BRNN 层中采用了 LSTM 细胞来学习远距离和闭合序列内的依赖性。网络输入包括氨基酸的 7 种代表性理化性质(Physio-chemical properties, PP),PSI-BLAST 的 20 维位置特异性评分矩阵(PSSM),以及来自 HHBlits 的 30 维隐马尔可夫模型特征。该方法数据集包含 5 789 个蛋白质,序列相似性截断值为 25%,X 射线分辨率低于 2.0 个 Å。从所有数据中,随机选择4 590种蛋白质作为训练集(TR4590),其余 1 199 用作独立测试集(TS1199)。文献[4]中指出捕获序列的长距离相互作用可以使三态二级结构预测准确率达 84%。

Spider3 提供单条蛋白质和批量蛋白质序列预测的功能,同时网站也提供了预测软件下载的功能。在线提交测试序列过程中,由于服务器资源有限,同一个 IP 和邮箱下提交序列总数不可超过 100 条,注意提交序列过程中序列不要换行。Spider3 网址为:http://sparks-lab.org/server/SPIDER3/。提交界面如图 3 所示。提交成功后,可在邮箱接收到最终结果,或者在网页端进行查看。

2.2 SPOT-1D

SPOT-1D 是目前较新的一种蛋白质二级结构预测服务器。作为 Spider3 的改进方法,SPOT-1D 在使用了双向递归神经网络的长短时记忆细胞(Long-Short-term memory Cells in Bidirectional recurrent neural networks, LSTM-BRNNs)基础上,结合了残余卷积网络(Residual Convolutional Networks, ResNets)[5],用来识别和传播整个序列中的短期和长期依赖关系,预测结果准确率得到了明显的提升,网络模型的描述在文献[5]的补充部分有详细的说明。该模型的特征输入由氨基酸的 7 种代表性理化

性质,SPOT-Contact 的预测接触图信息,PSSM 和隐马尔科夫模型特征组成,共 57 维特征输入。相比于 Spider3,SPOT-1D 的预测更加准确,除了模型的改进,SPOT-1D 从 PISCES 服务器中选取了更多数量的 10 029条蛋白质进行训练。使用界面和操作方法同 Spider3,但每次提交序列不可超过 5 条。SPOT-1D 的网址为:http://sparks-lab.org/jack/server/SPOT-1D/。

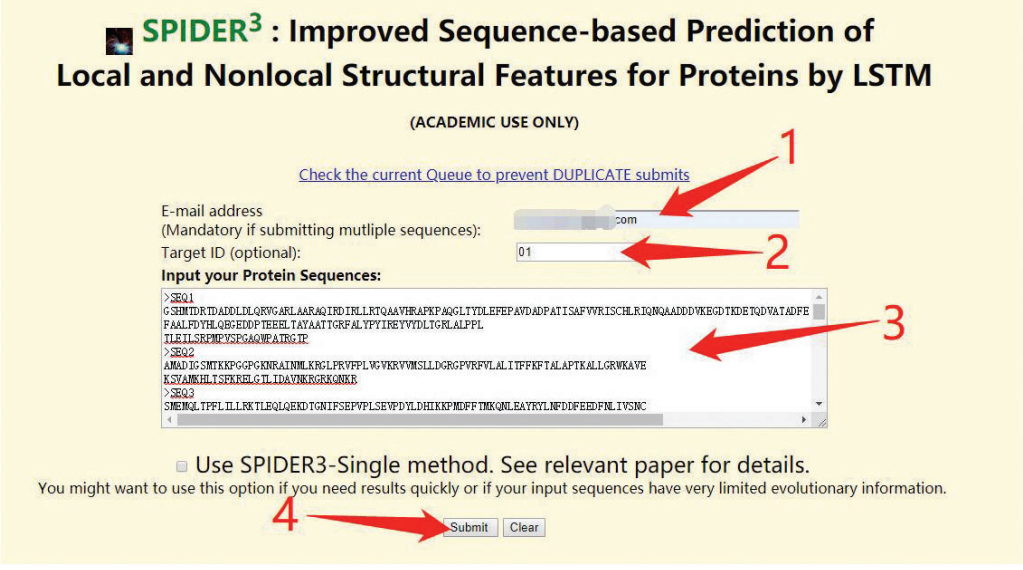

图 3　Spider3 提交测试序列

Fig.3　Submission test sequence in Spider3

注:首先,用户在图中标注 1 处输入预留邮箱,在标注 2 处输入工作名称,然后在下方标注 3 处根据图中示例格式输入序列,最后点击标注 4 处的“Submit”完成序列提交.

2.3　RaptorX

RaptorX 采用了名为深度卷积神经场(Deep convolutional neural fields, deepCNF)[6,11]的预测方法,该方法是深度卷积神经网络(Deep convolutional neural networks, DCNN)和条件神经场(Conditional neural fields, CNF)相结合。它能以分层的方式对复杂序列的结构关系进行建模,而且可以根据相邻残基之间的相关性建模。在 DeepCNF 中使用 DCNN 替换 CNF 中的浅层神经网络,以便捕获输入维度和输出标签之间的复杂关系,特别是对于在 PDB 中没有紧密同源性或具有稀疏序列谱的蛋白质具有很好的预测效果。针对紊乱蛋白质序列的预测,RaptorX 在网络中增加了 ROC 曲线下面积最大化(Area under the ROC Curve , AUC)方法训练[12]。该网络的特征输入由 21 维 PSSM 和具有 21 个元素的二进制向量(表示第 i 个位置上的氨基酸)组成,共 42 维。RaptorX 使用了 CullPDB 中 5 600 条蛋白质用作训练。该网站提供了批量预测的功能,提交方式如图 4 所示。在“My Jobs”里输入测试时提交的邮箱,等待结果链接。在线服务网址为:http://raptorx.uchicago.edu/StructurePrediction/predict/。

2.4　MUFOLD

MUFOLD 采用名为 Deep3I 的网络(Deep inception-inside-inception networks, Deep3I)[7]进行蛋白质二级结构预测。Deep 3I 由两个嵌套的可进行卷积操作的初始模块、卷积以及完全联通的致密层组成,有效地处理了氨基酸之间的局部和全局相互作用。MUFOLD 对训练集输入特征有非常细致的设计,训练集为由氨基酸理化性质,PSI-BLAST 特征和 HHBlits 特征组成的维度为 58 的特征向量。随机选取了 CullPDB 中的 9 000 条蛋白质用作训练集。该团队同时也利用初始胶囊网络的深度神经网络(Inception capsule networks)改善蛋白质 γ-转角预测[13]。测试过程如图 5 所示:输入邮箱和项目名称后,在下方勾选“Secondary Structure (3-states and 8-states)”,然后提交蛋白质序列,不允许序列字符断开或换行,且最多允许提交 10 个序列,每条序列的长度范围为 30 到 700。该服务器网址为:http://mufold.org/mufold-ss-angle/。

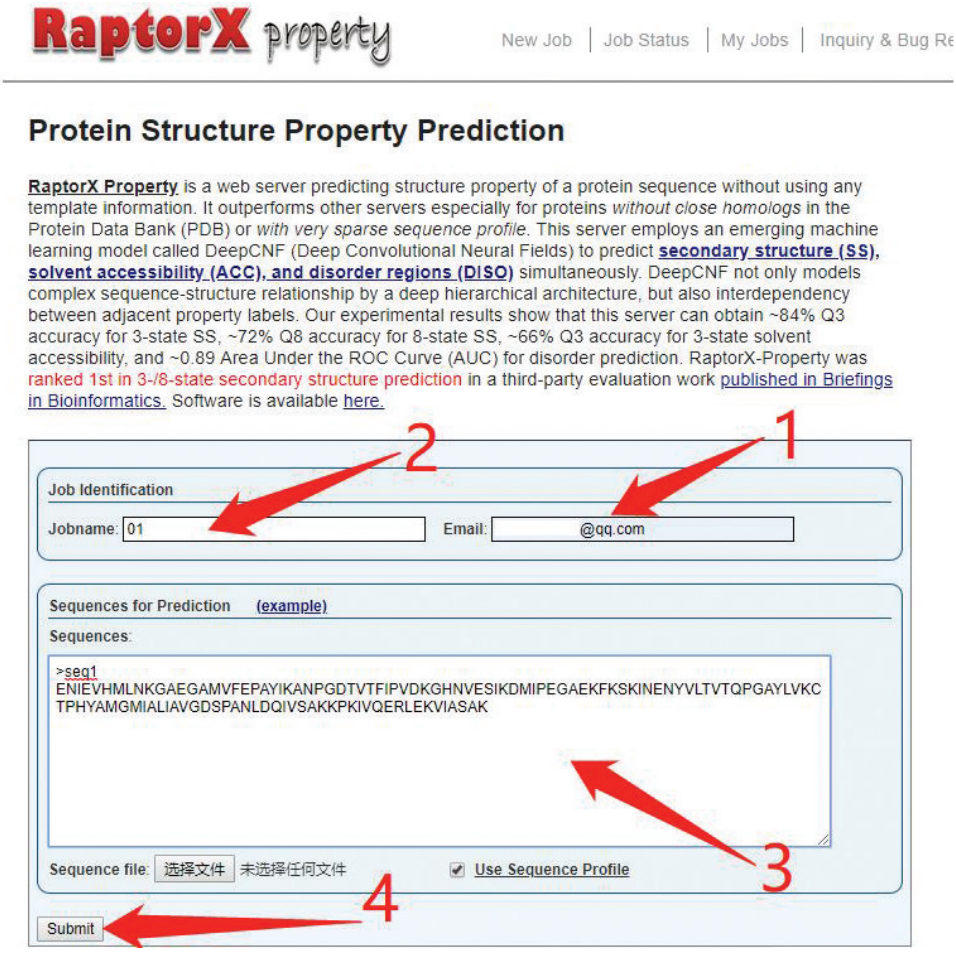

图 4 RaptorX 批量提交测试序列

Fig.4 Batch submission test sequence in RaptorX

注:首先,用户在图中标注 1 处输入预留邮箱,在标注 2 处输入工作名称,然后在下方标注 3 处根据图中示例格式输入序列,最后点击标注 4 处的“Submit”完成序列提交.

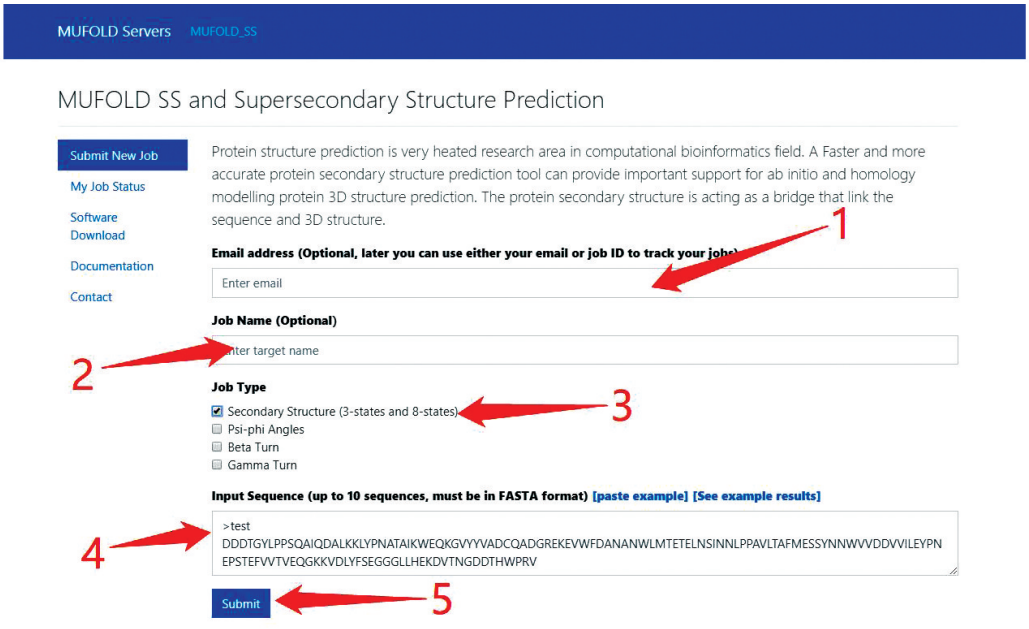

图 5 MUFOLD 提交测试序列

Fig.5 Submission test sequence in MUFOLD

注:首先,用户在图中标注 1 处输入预留邮箱,在标注 2 处输入工作名称,然后在下方标注 3 处选择“Secondary Structure (3-states and 8-states)”,然后在标注 4 处的文本栏中输入提交序列,最后点击下方“Submit”完成序列提交.

2.5 Psipred

Psipred 是常用的一种蛋白质二级结构预测服务器,该服务器聚合了多种蛋白质注释工具,提供分析方法作为软件下载。例如提供了序列和结构注释方法: Psipred, GenTHREADER, pGENTHREADER 等。在网络结构方面,Psipred 采用了两层前馈神经网络的体系,经交叉验证对网络性能进行评估。网络的输入是来自 PSI-BLAST 的 20 维特征矩阵。预测蛋白质二级结构的使用方法为:选择好所使用二级结构预测服务,然后输入序列,同样需要注意序列字符串不要换行,最后在输入的邮箱中接收结果。如图 6 所示。该服务网址为:http://bioinf.cs.ucl.ac.uk/psipred/。

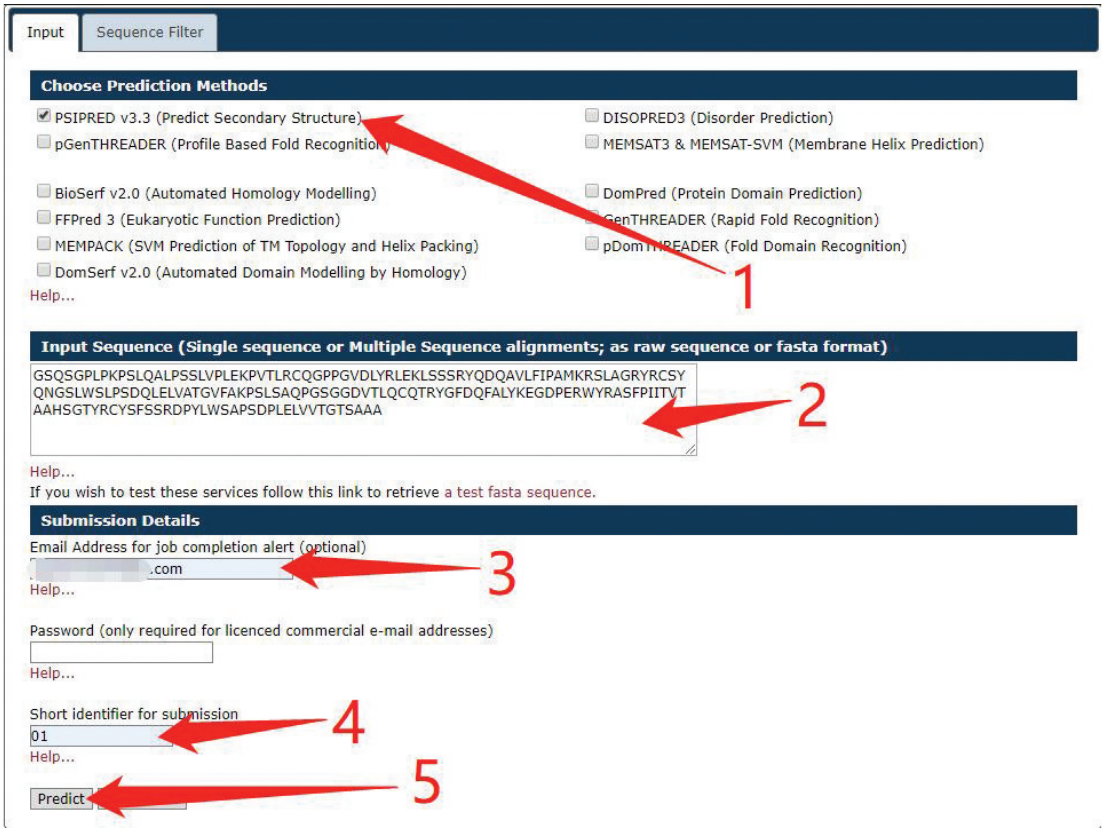

图 6　Psipred 批量提交测试序列

Fig.6　Batch submission test sequence in Psipred

注:首先,用户在图中标注 1 处选择服务“PSIPRED V3.3”,在标注 2 处的文本栏中输入提交序列, 在标注 3 和 4 处分别输入预留邮箱和工作名称,最后点击下方“predict”完成序列提交.

2.6 Jpred4

Jpred4 通过 JNet[8] 算法提供预测。在上个版本 JPred3[14] 中使用 JNet2.0 对蛋白质序列进行预测, JNet2.0 不使用频率文件,以 PSSM 和隐马尔科夫特征作为输入,使用两层来自 SNNS 神经网络包的人工神经网络,将隐藏层单元从 9 增加到 100。Jpred4 则基于 JNet2.0 神经网络的预测器进行了重新训练,通过使用1 358个 SCOPe/ASTRAL v.2.04 超家族域序列中的每个序列的一个代表进行 7 倍交叉验证来制作 JNet2.3.1,通过搜索 UniRef90v.2014_07 生成 PSI-BLAST 构建了每个序列的多重比对。除了对 JNet2.0 重新训练之外,JNet 中的 HMM 构建步骤已更新为 HMMer3。Jpred4 最终在 150 个未用于训练的超家族序列的盲测中评估其准确性,Q3 准确率可达到 82%。

该网站提供了批量预测的功能,如果只是提交单个序列则要在“Advanced options”中的“Select type of input”选项中,选中“Single Sequence”的“Raw/Fasta”模式;如果是批量在网页中输入蛋白质序列,则需要在“Single Sequence”下选中“Batch Mode”模式,然后输入接收结果的邮箱以及项目名(其中命名方式只可以是由字母数字和“_”字符组成)。批量提交过程如图 7 所示。最终结果将会发送到邮箱中,也可以在网页端等待查看。Jpred4 提供服务的网址为: http://www.compbio.dundee.ac.uk/jpred4/index.html。

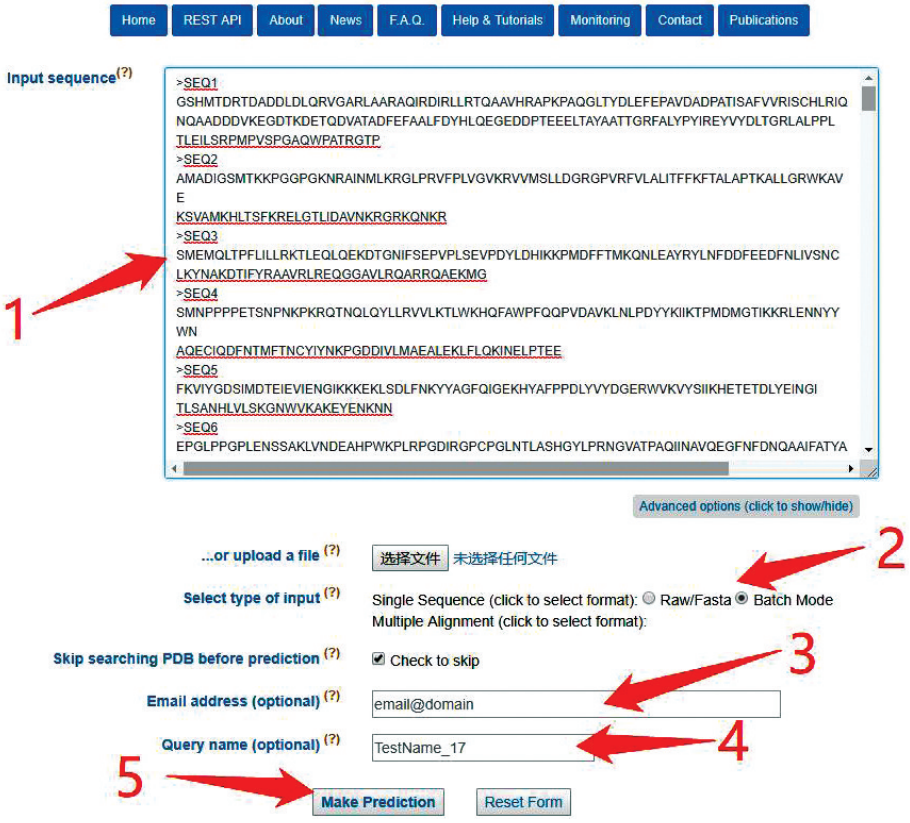

图 7 Jpred4 批量提交测试序列

Fig.7 Batch submission test sequence in Jpred4

注:用户首先在标注 1 处输入需要提交的序列,如果同图中一样提交批量测试,则在标注 2 处选择“Single Sequence”下的“Batch Mode”模式;如果是单个序列提交,则选择“Raw/Fasta”模式,然后在标注 3 和 4 处分别输入预留邮箱和工作名称,最后在标注 5 处点击“Make Prediction”完成工作提交.

3 结果评估

对上述七种服务器进行了预测结果评估,为保证实验数据量和公平性,测试集选取了 PDB 中 2018 年 8、9、10、11 月份发布的蛋白质[3,5,7],从中随机选取了 60 条 30%同源性,60 条 40%同源性和 60 条 70%同源性的蛋白质分别进行实验,最后又做出了这 180 条蛋白质的整体评估结果。实验数据集见表 1。

评估采用了七种衡量标准,分别为 Q3[2-14],Sov[2-6],边界识别率[3],内部识别率[3]和 C、E、H 每种独立结构识别率[4-5]的衡量标准。

3.1 Q3

在 8 态 DSSP[15]中,根据“G、H、I 转为 H(螺旋)”,“B、E 转为 E(折叠)”,“其他结构转为 C(转角)”将 8 态转为 3 态结构。Q3 为正确预测的氨基酸数占所有氨基酸的比例,计算公式如下:

$$Q3 = \frac{Q_C + Q_E + Q_H}{S} \times 100\%$$

其中,Q_C 为正确预测的转角数,Q_E 为正确预测的折叠数,Q_H 为正确预测的螺旋数,S 为总的氨基酸数。

3.2 Sov

Sov 是一种基于重叠片段比值的度量方式,设观测到的所有结构片段标记为 S_{ab},所有预测到的片段则标记为 S_{pr},而 S_a 是 S_{ab} 和 S_{pr} 状态相同的片段。任何观测到的残基长度被定义为 $\text{length}(S_{ab})$,对于 S_a 中任意一对片段,实际长度为 $\text{minov}(S_{ab}, S_{pr})$,至少有一个残基的长度总限度为 $\text{maxov}(S_{ab}, S_{pr})$。基于以上定义,Sov 的计算公式如下:

$$Sov = \frac{100}{n_{Saw}} \sum_{Sa} \left[\frac{\text{minov}(S_{ab}, S_{pr}) + \sigma(S_{ab}, S_{pr})}{\text{maxov}(S_{ab}, S_{pr})} \text{length}(S_{ab}) \right]$$

其中增设因子 $\sigma(S_{ab}, S_{pr})$,允许蛋白质结构中的观测片段边界处的变化,其定义为:

$$\sigma(S_{ob}, S_{pr}) = \min \begin{cases} (\text{maxov}(S_{ob}, S_{pr}) - \text{minov}(S_{ob}, S_{pr})) \\ \text{minov}(S_{ob}, S_{pr}) \\ \text{int}[\text{len}(S_{ob})]/2 \\ \text{int}[\text{len}(S_{pr})]/2 \end{cases}$$

int 为取整操作,例如 int [5.25]=5,N_{Sov} 为 S_a 中所有重叠片段中观察到的残基数量总和加上没有和预测状态相同的 S_{ab} 片段。

表 1　180 条数据集

Table 1　180 data set

Homology/%	Protein name									
30	5Y3S_A	5Y4U_A	5XZ4_A	5YAV_A	5MTW_C	5T2Q_A	5V85_A	5YCT_A	5YPS_A	5Y3S_A
	5Y42_A	5Y53_A	5Y24_A	5YBC_C	5OW2_A	5VWH_A	5VY2_A	5YDV_A	5YPV_A	5Y42_A
	5Y4M_A	5Y57_A	5Y6Y_B	5YBX_A	5OWM_A	5W2J_A	5W5T_A	5YDW_A	5YQ4_A	5Y4M_A
	5Y4N_A	5Y6L_A	5Y9R_A	5YCD_A	5OXZ_A	5W9A_A	5XLN_A	5YO2_A	5YRN_A	5Y4N_A
	5Y4Q_A	5Y6U_A	5Y9Z_A	5YCE_A	5OY1_A	5OGR_A	5XYS_A	5YOX_A	5YRY_A	5Y4Q_A
	5Y4T_A	5Y7D_A	5YAM_A	5YCL_A	5QGF_A	5UXY_A	5YCA_A	5YPD_A	5YS4_A	5Y4T_A
40	5X2I_A	5XWX_A	5Y2I_B	5Y6I_A	5YD7_A	5YGI_A	5YPP_A	5OWE_A	5Y8X_A	5Z51_A
	5OLD_A	5XX0_A	5Y2P_A	5Y7P_A	5YDD_A	5YI6_A	5YQ8_A	5WH9_A	5YLB_A	5ZCM_A
	5XNC_A	5XXJ_A	5Y2Q_D	5Y7X_A	5YED_A	5YIG_A	5YQP_A	5WJI_A	5YLN_A	5ZHR_A
	5XNE_A	5XXQ_A	5Y5B_A	5OL3_A	5YEJ_A	5YKA_A	5YSC_A	5WPH_A	6A5Y_A	5ZIH_A
	5XOB_Z	5XYB_A	5Y5C_A	5Y9B_A	5YFY_A	5YKR_A	5YSN_A	5XPJ_A	5OD7_A	5ZL6_A
	5XQZ_A	6DR7_A	5Y5M_A	5YCP_A	5YGG_A	5YL9_A	5YU2_A	5XTU_A	5Z0Q_C	5ZLV_F
70	5WOE_F	5OK2_A	5OU7_A	5ULB_A	5YEZ_A	5YIO_A	5YKJ_A	6A7V_A	5W7Z_A	5YUQ_A
	5WOV_A	5ONQ_A	5OUC_A	5V1V_A	5YGP_A	5YJA_A	5YKP_A	6A42_A	5YLW_A	5YUS_A
	5WOW_A	5OME_A	5OUF_A	5YAA_A	5YH4_A	5YJC_A	5YKU_A	6AGQ_A	5YT1_A	5YVK_A
	5WP1_A	5OS9_A	5OUR_A	5YBY_A	5YHK_A	5YJD_A	5YKZ_A	6AI4_A	5YT3_A	5YVN_A
	5WYN_A	5OSR_A	5WFV_A	5YBZ_A	5YHV_A	5YJO_A	6A4R_A	6AIX_A	5YTA_A	5YW3_A
	5OO3_A	5OTN_A	5WFY_A	5YC6_C	5YII_A	6BUB_A	6A7H_A	6DK0_A	5YTQ_A	5YX6_A

3.3　边界识别率和内部识别率

假设在一条长度为 N 的蛋白质序列中,第 n(1<n<N)个位置的结构同第 $n-1$ 和 $n+1$ 位置上的结构有一处不同,则称该处为边界。如果它前后位置的结构都与它相同,则称为内部[3]。

根据蛋白质同源性分类的所有服务器 Q3、Sov、边界准确率和内部准确率见表 2~表 4,180 条蛋白质的各项预测平均值见表 5。

从表 2 可以看出,同源性 30%的蛋白质数据集中,PSRSM 在 Q3、边界识别率和内部识别率上取得了最好的结果,分别达到了 89.49%,84.25% 和 90.91%,并且对转角 C 和折叠 E 的识别率也是最好的,准确率分别达到了 87.19% 和 90.27%。而 SPOT-1D 在 Sov 和螺旋 H 的识别率上结果要比 PSRSM 好一些,分别为 83.16%和 91.36%。

表 2　30%同源性数据集

Table 2　30% homology data set　　%

Servers	Q3	SOV	Boundary accuracy	Internal accuracy	Q_C	Q_E	Q_H
PSRSM	**89.49**	82.73	84.25	**90.91**	**87.19**	**90.27**	91.11
SPOT-1	87.58	**83.16**	79.27	89.68	84.80	84.91	**91.36**
MUFOLD	86.11	79.95	78.41	88.17	82.86	82.41	89.22
SPIDER3	84.94	80.55	74.28	87.89	84.52	75.50	85.79
RaptorX	82.95	78.11	71.82	86.30	85.38	76.70	79.08
Psipred	79.02	72.08	66.12	82.72	86.43	56.67	75.16
Jpred4	78.17	72.65	64.52	81.80	82.64	67.56	75.02

表 3 里 40%同源性的数据下,PSRSM 各项指标均为最好的结果,分别为 Q3:90.53%;Sov:84.71%;边界识别率:85.24%;内部识别率:91.25%;转角 C:87.34%;折叠 E:88.46%;螺旋 H:92.91%。SPOT-1D 紧随其后,Q3 为 88.52%,相差 2.01%,但 Sov 的表现依旧很出色,比 PSRSM 低约 0.5%。

在表 4 中,对于 70%同源性的蛋白质,PSRSM 除了内部识别率,其他指标均取得了最好的结果,分别为: Q3: 89. 87%; Sov: 86. 12%; 边界识别率: 83.65%;转角 C:89.08%,折叠 E:88.34%和螺旋 H: 89.64%。SPOT-1D 的内部识别率为 91.46%,其他指标同 PSRSM 的差距和在 40%同源性数据集的结果没有太大差别,约低 1%~2%。

表 5 为全部数据集的评估结果,PSRSM 各项指标全部取得了最好的结果: Q3: 89. 96%; Sov: 84.52%;边界识别率:84.37%;内部识别率:91.18%; 转角 C:87.88%,折叠 E:88.98%和螺旋 H:91.25%。

表 3 40%同源性数据集

Table 3 40% homology data set %

Servers	Q3	SOV	Boundary accuracy	Internal accuracy	Q_C	Q_E	Q_H
PSRSM	**90.53**	**84.71**	**85.24**	**91.25**	**87.34**	**88.46**	**92.91**
SPOT-1	88.52	84.27	78.70	90.65	84.22	83.86	90.52
MUFOLD	87.75	82.93	78.50	89.90	83.64	84.58	90.35
SPIDER3	85.90	81.97	74.34	88.78	83.65	76.36	88.31
RaptorX	84.09	79.53	71.38	87.62	84.57	77.69	78.74
Psipred	80.76	75.42	66.58	84.58	86.50	62.20	73.75
Jpred4	78.66	73.67	63.78	82.29	82.91	64.18	73.12

表 4 70%同源性数据集

Table 4 70% homology data set %

Servers	Q3	SOV	Boundary accuracy	Internal accuracy	Q_C	Q_E	Q_H
PSRSM	**89.87**	**86.12**	**83.65**	91.40	**89.08**	**88.34**	**89.64**
SPOT-1	88.44	85.12	78.36	**91.46**	87.09	85.70	86.74
MUFOLD	86.38	83.32	77.19	89.23	84.35	82.74	86.43
SPIDER3	86.47	82.81	74.85	90.54	86.78	79.36	84.13
RaptorX	83.78	80.05	70.76	88.14	86.23	79.81	72.13
Psipred	79.40	74.94	65.31	84.38	87.70	65.42	68.15
Jpred4	78.06	74.57	62.74	82.70	84.30	69.80	66.52

表 5 180 条数据集

Table 5 180 data set %

Servers	Q3	SOV	Boundary accuracy	Internal accuracy	Q_C	Q_E	Q_H
PSRSM	**89.96**	**84.52**	**84.37**	**91.18**	**87.88**	**88.98**	**91.25**
SPOT-1	88.18	84.19	78.78	90.60	85.37	84.85	89.59
MUFOLD	86.74	82.05	78.03	89.09	83.61	83.23	88.72
SPIDER3	85.77	81.78	74.49	89.07	84.99	77.18	86.14
RaptorX	83.61	79.23	71.32	87.35	85.39	78.14	76.78
Psipred	79.72	74.12	66.01	83.88	86.86	61.59	72.51
Jpred4	78.29	73.62	63.68	82.26	83.29	67.33	71.65

为了更加直观的对评估结果进行观察,将所有网站的 Q3 结果根据蛋白质长度做出了散点图,所选 180 条数据集中,蛋白质的长度范围为 34-552,如图 8 所示。可以看出 PSRSM(黄色)相对于其他颜色的位置更偏向于顶部,大部分服务器的预测准确率在 70%~90%,PSRSM 结果是优于其他服务器的。

表 6 对各服务器的预测方法、训练集、模型输入特征,Q3 准确率和使用效率方面做了总结。本此测试从 PDB 中随机选取了一条长度为 235 的蛋白质: 5XNE_A,测试各服务器从提交序列到获得结果的时间,结果为 Jpred4 最快,用时51 s;SPOT-1D 所需时间最长,为16 m 42 s。然后又随机选取了五条蛋白质:5WOV_A(长度 34)、5YIO_A(长度 121)、5YKU_A(长度 125)、5YVK_A(长度 225)、5Y5B_A(长度

228)做进一步的测试,结果为 Jpred4 用时最短,为 2 m 25 s,SPOT-1D 用时最长,为27 m 45 s。在 WEB 使用体验上,PSRSM、Spider3、RaptorX、Jpred4 均提供了批量测试的功能;除了 PSRSM,其他方法也提供了支持不同操作系统环境的软件下载服务。

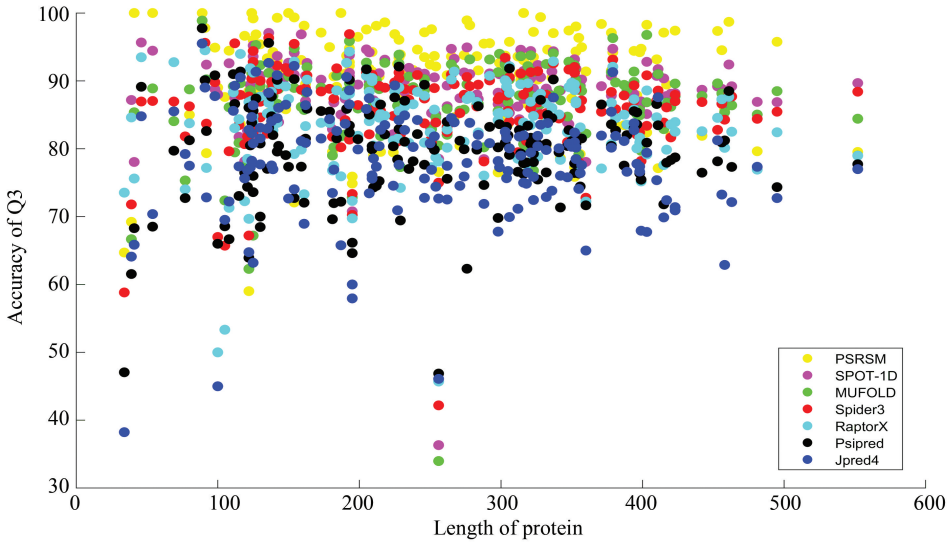

图 8　所有服务器的 Q3 散点图

Fig.8　Q3 scatter plot for all servers

表 6　各服务器方法总结

Table 6　Summary of methods for each server

Servers	PSRSM	Spider3	SPOT-1D	RaptorX	MUFOLD	Psipred	Jpred4
方法原理	分段特征提取+SVMs	LSTM+BRNNs	ResNets+LSTM-BRNNs	DCNN+CNF	CNN	前馈神经网络	SNNS 神经网络包
训练集数量	(ASTRAL+CullPDB) 15 696	(PISCES 服务器中选取) 4 590	(PISCES 服务器中选取)10029	(CullPDB) 5 600	(CullPDB) 9 000	——	(SCOPe/ASTRAL) 1 358
特征输入	PSSM	氨基酸 7 种理化性质+PSSM+隐马尔可夫模型特征	氨基酸 7 种理化性质+SPOT-Contact 预测接触图信息+PSSM+隐马尔可夫模型特征	PSSM+21 个元素的二进制向量	氨基酸理化性质+PSSM+隐马尔可夫模型特征	PSSM	PSSM+隐马尔可夫模型特征
Q3/%	89.96	85.77	88.18	83.61	86.74	79.72	78.29
是否支持批量预测	是	是	是	是	否	否	是
一条蛋白质(长度 235)运算时间	5 m 39 s	3 m 40 s	16 m 42 s	3 m 19 s	2 m 30 s	1 m 46 s	51 s
五条蛋白质运算时间	17 m 7 s	12 m 39 s	27 m 45 s	7 m 5 s	9 m 20 s	6 m 30 s	2 m 25 s

4　结　论

对 PSRSM、Spider3、SPOT-1D、RaptorX、MUFOLD,Psipred 和 Jpred4 七种在线服务的蛋白质二级结构预测效果进行了评估。整体来看,在多种比对方法下,PSRSM 绝大多数指标都取得了最优的结果。从方法选择角度来看,PSRSM 根据蛋白质长度划分不同子集和基于大数据集的训练方式,明显有较好的成效,而紧随其后的 SPOT-1D 多种深度学习方法和

大数据集的训练结合,Sov 的准确率也是非常稳定,效果出色。可以看出,蛋白质二级结构预测可以从结合多种深度学习方法,运用大数据进行模型训练做进一步的研究。

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]泽瓦勒贝 M, 鲍姆 J.O. 理解生物信息学[M]. 李亦学, 郝沛, 译. 北京: 科学出版社, 2012.
ZVELEBIL M, BAUM J O. Understanding bioinformatics [M]. LI Yixue, HAO Pei, Trans. Beijing: Science Press, 2012.

[3]MA Yuming, LIU Yihui, CHENG Jinyong. Protein secondary structure prediction based on data partition and semi-random subspace method[J]. Scientific Reports,2018,8(1): 9856. DOI:10.1038/s41598-018-28084-8.

[4]HEFFERNAN R, YANG Y, KULDIP P, 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.

[5]HANSON J, PALIWAL K, LITFIN T, et al. Improving prediction of protein secondary structure and solvent accessibility by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks [J]. Bioinformatics,2019,35(14):2403-2410. DOI:10.1093/bioinformatics/bty1006.

[6]WANG Sheng, PENG Jian, MA Jianzhu, et al. Protein secondary structure prediction using deep convolutional neural fields[J]. Scientific Reports, 2016, 6: 18962. DOI: 10.1038/srep18962.

[7]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.

[8]DROZDETSKIY A, COLE C, PROCTER J, et al. JPred4: A protein secondary structure prediction server[J]. Nucleic Acids Research,2015,43(1):389-394. DOI:10.1093/nar/gkv332.

[9]JONES D. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology,1999,292(2):195-202. DOI:10.1006/jmbi.1999.3091.

[10]朱树平, 刘毅慧. 蛋白质二级结构在线服务器预测评估[J]. 生物信息学,2019,17(1):53-60. DOI:10.12113/j.issn.1672-5565.201808002.
ZHU Shuping, LIU Yihui. Protein secondary structure online server predictive evaluation[J]. Chinese Journal of Bioinformatics,2019, 17(1): 53-60. DOI: 10.12113/j.issn.1672-5565.201808002.

[11]WANG Sheng, LI Wei, LIU Shiwang, 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.

[12]WANG Sheng, MA Jianzhu, XU Jinbo. AUCpreD: Proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields[J]. Bioinformatics,2016, 32: 672-679. DOI:10.1093/bioinformatics/btw446.

[13]FANG Chao, SHANG Yi, XU Dong. Improving protein gamma-turn prediction using inception capsule networks [J]. Scientific Reports, 2018, 8(1): 15741. DOI: 10.1038/s41598-018-34114-2.

[14]COLE C, BARBER J, BARTON G. The Jpred 3 secondary structure prediction server[J]. Nucleic Acids Research, 2008,36(suppl 2):197-201. DOI:10.1093/nar/gkn238.

[15]WAKAMURA K, HIROKAWA K, OPITA 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.

[责任编辑:吴永英]