DOI:10.16356/j.1005-2615.2018.05.006

基于加权朴素贝叶斯分类器和极端随机树的蛋白质接触图预测

金康荣 於东军

(南京理工大学计算机科学与工程学院,南京,210094)

摘要:提出一个新的基于集成学习的预测器(TargetPCM),对蛋白质接触图(特别是中长程)进行高精度的预测。首先,TargetPCM 使用加权朴素贝叶斯分类器(Weighted Naïve Bayes classifier,WNBC)融合 3 个接触图预测器的输出,其中 WNBC 中的权重参数通过粒子群算法优化得到;其次,将 WNBC 融合后的输出和基于序列的特征进行组合,得到更具鉴别能力的特征;在此基础上,应用极端随机树训练得到最终的蛋白质接触图预测模型。为了验证 TargetPCM 的有效性,在包含 98 个非冗余蛋白质的数据集上进行了测试。结果表明:对于短程、中程和长程接触,TargetPCM 的 Top $L/5$ 精度比现有最好的集成预测器(NeBcon)分别提高了 8.2%,16.1%和 5.3%。在 CASP11 上进一步的验证表明,对于短程、中程和长程接触,TargetPCM 的 Top $L/5$ 精度比现有最好的基于协同进化的集成预测器(MetaPSICOV)分别提高了 7.4%,9.1%和 7.5%。实验结果验证了本文所提蛋白质接触图预测方法的有效性。

关键词:模式识别与智能系统;蛋白质接触图;特征提取;加权朴素贝叶斯分类器;粒子群算法;极端随机树

中图分类号:TP391.4 **文献标志码**:A **文章编号**:1005-2615(2018)05-0619-10

Improved Contact Map Prediction Using Weighted Naïve Bayes Classifier and Extremely Randomized Trees

JIN Kangrong, *YU Dongjun*

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China)

Abstract: The accurate prediction of residue-residue contacts provides crucial help to the *ab initio* protein folding and 3D structure modeling, because the accurately predicted contacts can enforce useful constraints to the structure assembly. Recent CASP experiments have witnessed the prosperities on this topic and a number of promising protein contact map predictors have emerged in the past decades. Although much progress has been made, challenges (e.g., low prediction accuracy for long-range contacts) remain. Here we developed a new meta-based predictor, called TargetPCM, which can achieve high accuracy for protein contact map prediction. TargetPCM combines the outputs of three existing powerful contact map predictors by using a weighted Naïve Bayes classifier (WNBC), among which the weight parameters are optimized with particle swarm optimization (PSO) algorithm. Then, the outputs of WNBC are further combined with the intrinsic sequence-based features and fed to the final prediction model, which is trained with extremely randomized trees (ERT), for performing contact map prediction. Tested on 98 non-redundant proteins, our TargetPCM improves the Top $L/5$ accuracy over the

基金项目:国家自然科学基金(61373062,61772273)资助项目。

收稿日期:2017-11-01;**修订日期**:2017-12-28

通信作者:於东军,男,教授,博士生导师,E-mail:njyudj@njust.edu.cn。

引用格式:金康荣,於东军. 基于加权朴素贝叶斯分类器和极端随机树的蛋白质接触图预测[J]. 南京航空航天大学学报,2018,50(5):619-628. JIN Kangrong, YU Dongjun. Improved contact map prediction using weighted Naïve Bayes classifier and extremely randomized trees[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2018, 50(5):619-628.

best meta-based predictor (NeBcon) by 8.2%, 16.1% and 5.3%, respectively, for short-, medium- and long-range contacts. Further investigations on CASP 11 show that TargetPCM improves the Top L/5 accuracy over the best co-evolution based meta-server predictor (MetaPSICOV) by 7.4%, 9.1% and 7.5%, respectively, for short-, medium- and long-range contacts. Detailed analysis on the experimental results shows that both the effective utilization of complementary information from base predictors and the powerful learning capability of ERT account for the performance improvements of the proposed TargetPCM over existing contact map predictors.

Key words: pattern recognition and intelligent system; protein contact map; feature extraction; weighted Naïve Bayes classifier; particle swarm optimization; extremely randomized trees

蛋白质接触图(Protein contact map,PCM)包含了重要的蛋白质空间几何约束信息,有助于解决诸多蛋白质结构相关的生物信息学问题,如蛋白质折叠[1,2]、3D结构建模[3]和药物设计[4]等。使用传统的湿实验方法来测定蛋白质接触图非常耗时费力。因此,研发蛋白质接触图预测的自动化计算方法成为当下迫切的需求。近年来,用于蛋白质接触图预测的计算方法已经有了长足的发展[5,6]。

在蛋白质中,当两个残基的 C_β 原子(甘氨酸的情况为 C_α)之间的空间距离小于一定的阈值(如8 Å),则认为这两个残基是相互接触的[6,7]。一个蛋白质的接触图通常可以用一个对称矩阵来表示,矩阵中各元素的值为0或1,分别对应于非接触对或接触对。根据残基间的序列间距的不同,残基接触可以划分为短程、中程和长程3种类型之一。短程、中程和长程序列间距范围的标准分别为6~11,12~23和≥24,该标准已经广泛应用于包括CASP(Critical assessment of techniques for protein structure prediction)竞赛等相关领域[6,7]。

在3种类型的残基接触中,中长程接触对于蛋白质的折叠起着至关重要的作用,因而引起了人们的广泛关注[8]。遗憾的是,虽然蛋白质接触预测取得了重大进展,但对于中长程接触的预测精度仍不尽如人意,远不能满足蛋白质三维结构建模等实际应用的需要[5]。在最近的CASP竞赛中(CASP 9和CASP 10),对于那些自由建模的蛋白质预测对象,即使是那些最先进的方法,长程接触预测的Top L/5的平均精度仍然低于13%[9,10]。因此,研发能够进行高精度中长程接触图预测的新方法就有着迫切需求。

现阶段,接触图预测方法可大致分为两类:基于模板的方法和基于序列的方法[6]。对于一个待预测的蛋白质,基于模板的方法使用结构已知的蛋白质模板进行同源建模或线程化来预测蛋白质的接触图;基于序列的方法则是利用序列衍生特征来进行接触图的预测[11,12]。基于模板的方法在很大程度上依赖于已知结构的模板的质量,当模板的质量欠缺时,往往不能得到令人满意的预测精度。因此,基于序列的方法受到越来越多的关注,特别是在海量蛋白质序列快速累积的后基因组时代。

基于序列的接触图预测方法又可以细分为两个子类,即基于协同进化的方法和基于机器学习的方法。基于协同进化的方法,比如PSICOV[13],GREMLIN[14],FreeContact[15]和CCMpred[16],均是基于接触残基共突变的假设,通过寻找多序列比对中的相关突变残基对来预测蛋白质残基接触;而基于机器学习的方法,比如BETACON[17],SVMcon[5],SVMSEQ[18],CMAPpro[19],PhyCMAP[20],Evfold[21]和NNcon[22],则是使用结构已知的蛋白质接触图作为训练数据,进而利用蛋白质序列特征训练预测模型来进行残基接触的预测工作。

基于协同进化的方法很大程度上依赖于从底层蛋白质数据库中获得的同源序列的数目和多样性。如果底层蛋白质数据库能够提供高质量的多序列比对(Multiple sequence alignments,MSA),则可以通过基于协同进化的方法实现对目标蛋白质序列进行可靠并且具有宽广覆盖域的接触图预测[13,23,24]。但是,如果目标序列获得的是质量较低的多序列比对,则不能得到令人满意的预测。基于机器学习的方法有较好的鲁棒性,不依赖于大量的多样化同源序列;但是,对于那些有大量同源序列的蛋白质,这些方法的预测精度往往要低于基于协同进化的方法。此外,因为训练样本里短程接触样本的数量远远大于长程接触样本,所以基于机器学习的方法预测出的接触图的覆盖域往往较窄[6,23]。

如上所述,基于协同进化和基于机器学习的方法各有其优缺点。因此,通过集成学习将两类方法进行融合,有望提升蛋白质接触图预测的精度。近年来,已经出现了一些融合多个基于协同进化的和(或)基于机器学习的集成方法。其中,最具有代表性的集成预测方法为MetaPSICOV[25]和R_2C[6]。最近,文献[26]提出了一个先进的集成方法NeBcon,该方法首先使用朴素贝叶斯分类器(Naïve Bayes classifier,NBC)融合8个现有的基于协同进化和基于机器学习的预测器,然后将NBC模型的输出与结构特征相结合,并在此基础上,应用一个

简单的反向传播网络训练得到最终的接触图预测模型。

虽然 NeBcon[26] 在接触图预测方面取得了显著的进展，但是仍有进一步提升的空间。首先，NeBcon 所融合的 8 个现有基础预测器之间可能存在冗余；其次，NeBcon 使用 NBC 进行融合，NBC 对每个基础预测器赋予相同的权重。而事实上，8 个基础预测器的预测质量可能有所不同，通过学习自动获得权重较为合理。因此，使用加权朴素贝叶斯分类器（Weighted Naïve Bayes classifier，WNBC）[27-29] 而不是 NBC 更符合实际情况；再次，可尝试其他先进的学习算法代替 BP 神经网络训练模型进一步提高预测性能。鉴于此，本文提出了 NeBcon 的增强版本，称为 TargetPCM。与 NeBcon 相比，TargetPCM 首先将 8 个基础预测器减为 3 个并采用加权朴素贝叶斯方法融合，而且使用粒子群算法（Particle swarm optimization，PSO）优化得到各分类器的权重系数[30-32]。然后，将 WNBC 的输出与基于序列的特征相结合形成新的特征向量。最后，应用极端随机树（Extremely randomized trees，ERT）训练得到最终的预测模型（TargetPCM）[33]。

1　实验材料与方法

1.1　基准数据集

本文采用 NeBcon[26] 等使用的训练数据集 Train_517。Train_517 中包含 517 个蛋白质序列，并且任意两条序列之间的同源性低于 25%。测试数据集分别采用 NeBcon[26] 使用的 Test_98 数据集及 CASP11 官网（http://predictioncenter.org/casp11/）中获得的测试数据集（下文中称为 CASP11 数据集）。Test_98 数据集由 98 个两两序列同源性小于 25% 的非冗余蛋白质组成。CASP11 数据集包括 103 个 CASP11 蛋白质，是在原始的 105 个 CASP11 的蛋白质基础上移除了 2 个残基数小于 50 的蛋白质（T0759-D1 和 T0820-D2）。103 个 CASP11 蛋白质中，有 48 个被认为是难预测的。原因在于：对于这 48 个蛋白质中的任意一个，参与竞赛的前 50% 的预测服务器预测出的 TM-score 平均得分低于 0.5[6]。

1.2　预测模型的结构框架

TargetPCM 使用基于粒子群优化算法的加权朴素贝叶斯方法（Weighted Naïve Bayes classifier based on particle swarm optimization，WNBC-PSO）融合 3 个互补的预测器（即基于机器学习的预测器（RFcon）、基于协同进化的预测器（PSICOV[13]）和基于集成学习的预测器（MetaPSICOV[25]）。特别地，RFcon 是本文通过复现 SVMcon[5] 并将其中的支持向量机算法[34] 替换成了随机森林算法[35] 得到的。

图 1 给出了接触图预测器 TargetPCM 的结构框图。对于 1 条蛋白质序列，首先使用 3 个基础预测器对其预测；接着通过 WNBC-PSO 融合 3 个基础预测器的输出从而得到基于预测器的特征，然后与从序列中提取的特征组合得到更有判别力的特征；在此基础上，应用 ERT 算法训练得到最终的预测模型来进行蛋白质接触图预测。

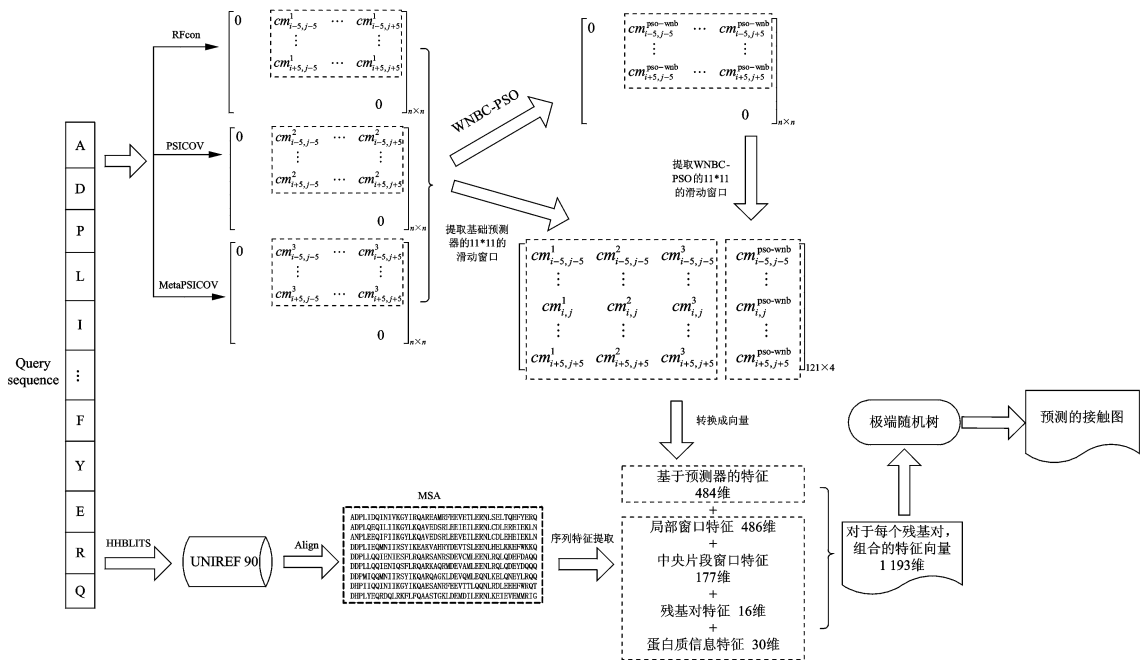

图 1　TargetPCM 结构框架

Fig. 1　The framework of TargetPCM

1.3 用 WNBC-PSO 提取基于预测器的特征

1.3.1 加权朴素贝叶斯分类器

由于朴素贝叶斯分类器(NBC)的条件独立性假设在实际情况中很难得到满足,朴素贝叶斯分类器认为所有属性对决策的分类重要性程度是相同的,即权重均为1。而事实上,不同属性对分类的影响也是有差别的。

WNBC[27-29]是传统的 NBC 的一个扩展。相对于 NBC,WNBC 考虑各个基础分类器的重要性程度赋予其不同的权重。

假设 N 是基础接触图预测器的数目,对于一条蛋白质查询序列中的每个残基对,可以得到 N 个预测得分,用 N 维向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n,\cdots,x_N)^{\mathrm{T}}$ 表示,其中 x_n 是由第 n 个基础预测器预测得到的接触得分。假设 $C=\{C_1,C_2,\cdots,C_m,\cdots,C_M\}$ 是 M 个类标签的集合。那么,残基对属于 C_m 类的后验概率公式为

$$P(C_m\mid\boldsymbol{X})=\frac{P(C_m)P(\boldsymbol{X}\mid C_m)}{P(\boldsymbol{X})}=\frac{P(C_m)\prod_{n=1}^{N}P(x_n\mid C_m)^{w_n}}{P(\boldsymbol{X})} \tag{1}$$

式中:$P(C_m)$是 C_m 类的先验概率;$P(\boldsymbol{X}|C_m)$是 C_m 类中 $\boldsymbol{X}$ 的条件概率;$P(\boldsymbol{X})$是证据因子,并且 w_n 是第 n 个基础预测器的权重系数。

在本文中,$C_m\in\{0,1\}$,其中1和0分别代表残基对是接触的或是不接触的。因此,残基对属于接触类的后验概率表示为

$$P(1\mid\boldsymbol{X})=\frac{P(1)P(\boldsymbol{X}\mid 1)}{P(\boldsymbol{X})}=\frac{P(1)\prod_{n=1}^{N}P(x_n\mid 1)^{w_n}}{P(1)\prod_{n=1}^{N}P(x_n\mid 1)^{w_n}+P(0)\prod_{n=1}^{N}P(x_n\mid 0)^{w_n}} \tag{2}$$

基于训练数据集 Train_517,可以使用 NeBcon[26]中描述的方法计算得到在式(2)中的4个参数,即 $P(0)$,$P(1)$,$P(x_n|1)$和 $P(x_n|0)$。另外,针对短程、中程和长程接触,需要分别计算这4个参数来加强模型的特异性[26]。

1.3.2 使用 PSO 算法优化 WNBC 的权重

本文使用 PSO 优化 WNBC 中每个基础分类器的权重[30-32,36,37]。实验中,1组基础接触图预测器的权重被编码成1个候选粒子,并且将 WNBC 在训练集 Train_517 上的 Top $L/5$ 的平均精度作为适应度函数。WNBC-PSO 的过程简述如下:

步骤1 设定粒子群的规模和最大迭代次数,并且随机初始化粒子群中每个粒子的位置。

步骤2 标准化每个粒子的位置作为基础分类器的权重,使用 WNBC 融合各个基础分类器并对数据集 Train_517 分类预测,计算得到每个粒子的适应度,来更新个体极值和全局极值,从而调整粒子群中每个粒子的位置和速度。

步骤3 重复上述过程直到达到最大迭代次数,迭代结束,找到最优粒子。

1.3.3 基于预测器的特征提取

对于1条蛋白质查询序列中的每个残基对(i,j),都可以得到1个4维向量,其中的4个元素分别是3个基础预测器(即 RFcon,PSICOV[13]和 MetaPSICOV[25])预测的接触得分和 WNBC-PSO 融合的后验概率。另外,为了提高特征的鲁棒性,设置了分别以残基 i 和残基 j 为中心的11个残基的滑动窗口[5,26]。这样,对于每个残基对(i,j),通过组合两个窗口中的所有残基对的四维向量可以得到1个484维($484=11\times11\times4$)的基于预测器的特征向量。

1.4 基于序列的特征提取

除了基于预测器的特征,本文对每个残基对(i,j)提取基于序列的特征。实验中,采用 SVMcon[5]中的方法提取基于序列的特征,其中包括486维的局部窗口特征,177维的中央片段窗口特征,16维的残基对特征以及30维的蛋白质信息特征,基于序列的特征的详细提取方法参见文献[5]。

1.5 训练 ERT 预测模型

本文组合使用了基于预测器的特征和基于序列的特征训练 ERT 预测模型。首先,将每个残基对的序列特征和预测器特征组合形成最终的1 193($=486+177+16+30+484$)维的特征向量。然后,基于这样鲁棒的特征表示,进一步应用极端随机树(ERT)[33]在训练集 Train_517 上训练预测模型。对于短程和中程接触类型,所有接触的残基对的特征向量组成正训练样本集,同时所有非接触的残基对的特征向量组成负训练样本集。对于长程接触,由于训练样本集过大,因此本文构建了一个100万的长程训练数据集,其中包括所有的长程接触样本以及部分随机采样的长程非接触样本。为了加强训练模型的特异性,对于短程、中程和长程接触,需要分别训练 ERT 模型(由 scikit-learn[38]实现)。

2 结果与讨论

本文通过与其他主流的接触图预测器的性能进行比较分析,系统地评估 TargetPCM 的性能。评价过程中,采用得分最高的前 K 个预测的精确度 Accuracy 作为评价指标,这也是本领域广泛应

用的评价指标。

对于 1 个包括 L 个残基的蛋白质，可得到 $(L+1-6)\times(L-6)/2$ 个残基对，其中 6 是在实验中所使用的序列间距的最小值。一个训练好的预测器首先预测每个残基对接触的概率得分然后对所有预测的得分进行降序排序。之后，使用 Accuracy$=N_{\mathrm{corr}}(K)/K$ 计算得分最高的前 K 个预测的 Accuracy 来评估预测器的性能，其中 $N_{\mathrm{corr}}(K)$ 是得分最高的前 K 个预测中正确预测的残基对数目。另外，由于真实接触的总数量与蛋白质长度近似成线性关系[39]，因此 L，$L/2$，$L/5$ 和 $L/10$ 是 K 的几个常用值。

2.1　WNBC-PSO 组合预测器性能分析

首先研究通过 WNBC-PSO 组合多个预测器的预测性能。在训练集 Train_517 上，分别使用朴素贝叶斯分类器(NBC)和基于粒子群算法的加权朴素贝叶斯分类器(WNBC-PSO)组合 3 个基础预测器(PSICOV[13]，RFcon 和 MetaPSICOV[25])。实验中，粒子群的规模和最大迭代次数分别设为 10 和 100，并针对短程、中程和长程接触分别单独优化 WNBC 的权重。另外，本文尝试增大粒子群规模或是增加最大迭代次数，发现 WNBC 的权重并无显著变化，反而增加了实验的额外开销，表明这两个实验参数的设置对于优化 WNBC 的权重是有效的。

在包括 98 个蛋白质序列的 Test_98 和 103 个蛋白质序列的 CASP11 上，对两个训练好的集成预测器(NBC 和 WNBC-PSO)进行测试，表 1 汇总了两个集成预测器和 3 个基础预测器的 Top $L/5$ 的预测精度。PSICOV 和 MetaPSICOV 的结果是将测试序列输入到从各自的网站上下载下来的训练模型中得到的；RFcon 的结果是将测试序列输入到在 SVMcon 的训练数据集上训练的模型中得到的；而 NBC 和 WNBC-PSO 的结果是将测试序列输入到在 Train_517 上训练的相应模型中得到的。

从表 1 中，可以得到如下两点结论：

(1)通过集成多个互补的预测器，接触图预测的精度得到进一步提高。和 3 个独立的基础预测器(RFcon，PSICOV 和 MetaPSICOV)比较，对于短程、中程和长程接触两个集成预测器(NBC 和 WNBC-PSO)都取得了显著的性能提升。分析取得性能提升的可能原因是集成预测器充分利用了独立的基础预测器所提供的互补信息，从而提升了整体的预测精度。

表 1　在 Test_98 和 CASP11 测试集上 3 个基础预测器以及两个集成预测器的 Top $L/5$ 的预测精度

Tab. 1　Accuracy of Top $L/5$ contact predictions by three base predictors and two combined predictors on the 98 targets in Test_98 and the 103 CASP11 targets

数据集	预测器	短程/%	中程/%	长程/%
Test_98	PSICOV[13]	28.63	31.06	37.79
	RFcon	55.52	46.85	28.76
	MetaPSICOV[25]	63.65	62.70	60.95
	NBC	66.78	64.22	61.00
	WNBC-PSO	66.84	64.71	63.71
CASP11	PSICOV[13]	21.67	25.55	34.21
	RFcon	48.09	40.30	24.54
	MetaPSICOV[25]	58.70	58.59	54.73
	NBC	59.50	59.33	53.06
	WNBC-PSO	60.02	60.19	56.59

(2)对于短程、中程和长程接触，NBC 算法的改进版 WNBC-PSO 均取得了最优的预测性能以及最高的精度，WNBC-PSO 在 Test_98 上的精度分别为 66.84%，64.71%和 63.71%，在 CASP11 上的精度为 60.02%，60.19%和 56.59%。WNBC-PSO 取得比 NBC 更优越性能的原因是 WNBC-PSO 使用 PSO 来学习 3 个基础预测器的权重，合理地赋予基础预测器恰当的权重，有助于区分基础预测器的重要性程度，从而加强整体的预测性能。而 NBC 算法赋予给不同的基础分类器的权重是相等的，即认为每个基础分类器对最终分类决策的影响是一致的，这种方式无法区分效果较差和效果较好的预测器对分类的重要性程度，从而降低整体的预测效果。

2.2　TargetPCM 预测模型性能分析

为了更进一步地提高接触图预测的性能，本文组合基于预测器的特征及基于序列的特征训练 ERT 预测模型，称为 TargetPCM。实验中两个重要参数，即生成树的数目(n_{Tree})和在每次分裂时随机抽取为候选的维数(m_{Try})都是分别设置为 500 和 60。表 2 在测试集 Test_98 和 CASP11 上比较了 WNBC-PSO 和 TargetPCM 之间的 Top $L/5$ 的预测精度。

如表 2 所示，在 Test_98 和 CASP11 测试集上，TargetPCM 的预测精度得到显著的提升。在 Test_98 上，对于短程、中程和长程接触 TargetPCM 取得的精度分别是 70.43%，66.65%和 66.11%，比 WNBC-PSO 的精度提升了 3.59%，1.94%和 2.40%。而且同样可以在 CASP11 上观察到相似的精度提升。分析 TargetPCM 优于 WNBC-PSO 的原因如下：

由于 TargetPCM 结合了从 WNBC-PSO 得到

的基于预测器的特征(后验概率得分)以及每个残基对的基于序列的特征。这个鲁棒的特征表示方法加强了用于接触图预测的残基对的特征判别能力,并且结合 ERT 模型强大的分类能力,因此 TargetPCM 可以取得更卓越的接触图预测性能。

表 2 在 Test_98 和 CASP11 测试集上,WNBC-PSO 和 TargetPCM 的 Top $L/5$ 的预测精度比较

Tab. 2 Comparison of accuracy of Top $L/5$ predictions for short-, medium-, and long-range contacts between WNBC-PSO and TargetPCM on Test _ 98 and CASP11

数据集	预测器	短程/%	中程/%	长程/%
Test_98	WNBC-PSO	66.84	64.71	63.71
	TargetPCM	70.43	66.65	66.11
CASP11	WNBC-PSO	60.02	60.19	56.59
	TargetPCM	63.06	63.91	58.81

2.3 与主流接触图预测器比较

2.3.1 在 Test_98 测试集上比较

为了进一步评估本文提出的 TargetPCM 的预测性能,本文将它与主流的接触图预测器比较,包括基于机器学习的预测器(RFcon,BETAcon[17],SVMcon[5]和 SVMSEQ[18]),基于协同进化的预测器(CCMpred[16],PSICOV[13] 和 FreeContact[15])以及基于集成学习的预测器(MetaPSICOV[25],NeBcon[26] 和 TargetPCM),比较结果如表 3 所示。SVMcon,SVMSEQ,BETAcon,FreeContact,PSICOV,CCMpred 和 MetaPSICOV 的结果是将测试序列输入到从各自网站下载的训练模型中得到的;NeBcon 的结果是摘自文献[26];RFcon 的结果是将测试序列输入到在 SVMcon 的训练数据集上训练的模型中得到的;TargetPCM 的结果是将测试序列输入到在 Train_517 上训练的相应模型中得到的。

从表 3 中可知:

(1)对于短程和中程的接触,基于机器学习的接触预测器比基于协同进化的预测器性能更好。

对于短程接触,4 个基于机器学习的预测器的平均精度是 51.86%,比 3 个基于协同进化的预测器的平均精度(即 30.11%)高 21.75%。

对于中程接触,4 个基于机器学习的预测器的平均精度比 3 个基于协同进化的预测器的平均精度高 8.21%。

(2)对于长程接触,基于机器学习的预测器的预测性能不如基于协同进化的预测器,4 个基于机器学习的预测器的平均精度仅仅只有 27.84%,明显低于 3 个基于协同进化的预测器的平均精度 42.29%。

表 3 在 Test_98 测试集上,TargetPCM 和其他主流的预测器的 Top $L/5$ 接触预测精度的比较

Tab. 3 Comparison of accuracy of Top $L/5$ contact predictions between TargetPCM and other existing predictors on Test_98

预测器类型	预测器	短程/%	中程/%	长程/%
机器学习	SVMcon[5]	45.52	40.54	25.86
	SVMSEQ[18]	51.24	43.75	25.63
	BETAcon[17]	55.16	44.12	31.12
	RFcon	55.52	46.85	28.76
协同进化	FreeContact[15]	28.09	35.36	41.00
	PSICOV[13]	28.63	31.06	37.79
	CCMpred[16]	33.60	40.41	48.08
集成学习	MetaPSICOV[25]	63.65	62.70	60.95
	NeBcon[26]	65.10	57.40	62.80
	TargetPCM	70.43	66.65	66.11

上述观察结果与现有研究[6,26]是相一致的。

(3)3 个集成预测器相比起基于机器学习和基于协同进化的预测器具有压倒性优势。对于短程、中程和长程接触,3 个集成预测器的平均精度分别达到了 66.39%,62.25%和 63.29%。这些结果显著表明了集成多个独立的和互补的预测器是进一步提高接触预测精度的一条有效途径。

(4)对于短程、中程和长程接触,TargetPCM 比 NeBcon 的预测精度均有显著提高。虽然 TargetPCM 和 NeBcon[26]有着相似的预测结构,但是在预测精度上,对于短程、中程和长程接触,TargetPCM 比 NeBcon 的预测精度分别提高了 8.2%(=(70.43−65.10)/65.10),16.1%(=(66.65−57.40)/57.40)和 5.3%(=(66.11−62.80)/62.80),TargetPCM 优于 NeBcon 的原因包括如下 5 个方面:

(1)TargetPCM 融合了 3 个互补的具有代表性的基础预测器来生成基于预测器的特征。NeBcon 融合了 8 个基础预测器,其中部分基础预测器的效果较差以及互相之间可能存在冗余,从而导致整体的预测性能降低。

(2)TargetPCM 采用 PSO 算法优化各分类器的权重。训练过程中,TargetPCM 为不同的分类器分配不同的基础权重,而 NeBcon 使用 NBC 组合 8 个基础预测器,并给每个预测器赋予相同的权重。显然,不同的基础预测器具有不同的预测质量,若分配相同的权重,会使得预测效果较差的预测器和较优的预测器对融合的结果具有相同的地位,从而降低整体的预测性能,因此分配适当的权重更为合理。TargetPCM 仔细考虑各个基础预测

器的重要性差别，所以使用更合理的 WNBC 融合基础预测器，并使用 PSO 算法优化每个基础预测器的权重，从而提高整体的预测性能。

(3)TargetPCM 包括的基于序列的特征达到了 709 维，虽然 NeBcon 所提取的基于序列的特征也有 596 维，但仍不如 TargetPCM。因此更广泛的序列信息使得 TargetPCM 在训练时能学习到更有效的规律，从而提高模型的分类性能。

(4)TargetPCM 含有的基于预测器的特征不仅仅包括加权朴素贝叶斯分类器的后验概率，而且还包括 3 个基础分类器（RFcon，PSICOV 和 MetaPSICOV）的接触得分，而 NeBcon 仅仅只有朴素贝叶斯分类器的后验概率。因此 TargetPCM 提取的基于预测器的特征达到了 484 维，而 NeBcon 只有 121 维，对预测器信息的有效利用使得 TargetPCM 具有更强大的分类判别能力。

(5)TargetPCM 基于极端随机树算法训练最终的预测模型。极端随机树算法是 1 种分类效果出色的集成学习算法，相对于 NeBcon 中使用的反向传播算法，更适合处理高维特征的样本数据。因此分类效果出色，而且能够实现并行化计算，这也是 TargetPCM 取得更优异性能的一个重要原因。

2.3.2　在 CASP11 测试集上比较

本文在 CASP11 测试集上与其他主流的接触图预测器进行比较，验证 TargetPCM 的优越性。在预测器的选择上，RFcon 和 PSICOV[13] 分别是基于机器学习和基于协同进化中具有代表性的两个预测器，R_2C[6] 和 MetaPSICOV[25] 是效果比较出类拔萃的集成预测器。对于 CASP11 中的 103 个蛋白质，NeBcon[26] 仅仅提供了带有自由建模域的 33 个蛋白质的预测结果，因此不包括它作比较。

表 4 汇总了在 CASP11 的 103 个蛋白质上，本文的 TargetPCM 和其他现有的预测器的 Top $L/5$ 预测精度的比较，表 5 汇总了在 CASP11 的 48 个难预测的蛋白质上与其他预测器的 Top $L/5$ 预测精度的比较。表 4 和表 5 中的结果是按照如下方式获得的：PSICOV 和 MetaPSICOV 的结果是将测试序列输入到从各自的网站上下载下来的训练模型中得到的；RFcon 的结果是将测试序列输入到在 SVMcon 的训练数据集上训练的模型中得到的；NBC 和 WNBC-PSO 的结果是将测试序列输入到在 Train_517 上训练的相应模型中得到的；R_2C 的结果是摘自文献[6]；TargetPCM 的结果是将测试序列输入到在 Train_517 上训练的相应模型中得到的。

表 4　在 CASP11 的 103 个蛋白质上，TargetPCM 和其他现有预测器的 Top $L/5$ 接触预测精度的比较

Tab. 4　Comparison of accuracy of Top $L/5$ contact predictions between TargetPCM and other existing predictors on the 103 targets in CASP11

预测器类型	预测器	短程/%	中程/%	长程/%
协同进化	PSICOV[13]	21.67	25.55	34.21
机器学习	RFcon	48.09	40.30	24.54
集成学习	R_2C[6]	48.50	41.90	37.60
	MetaPSICOV[25]	58.70	58.59	54.73
	TargetPCM	63.06	63.91	58.81

表 5　在 CASP11 的 48 个难预测的蛋白质上，TargetPCM 和其他现有预测器的 Top $L/5$ 接触预测精度的比较

Tab. 5　Comparison of accuracy of Top $L/5$ contact predictions between TargetPCM and other existing predictors on the 48 hard targets in CASP11

预测器类型	预测器	短程/%	中程/%	长程/%
协同进化	PSICOV[13]	12.92	12.89	16.28
机器学习	RFcon	47.71	41.29	19.07
集成学习	R_2C[6]	49.40	40.20	25.60
	MetaPSICOV[25]	54.67	50.69	34.31
	TargetPCM	57.78	56.40	38.92

从表 4 和表 5 中可知，本文提出的 TargetPCM 不仅仅远远优于 PSICOV[13] 和 RFcon，而且还领先于 MetaPSICOV[25] 和 R_2C[6]。在 CASP11 的 103 个蛋白质上，对于短程、中程和长程接触 TargetPCM 分别获得了 63.06%，63.91% 和 58.81% 的精度，即 TargetPCM 的 Top $L/5$ 精度比最好的基于协同进化的集成预测器（MetaPSICOV[25]）分别提高了 7.4%（=(63.06−58.70)/58.70），9.1%（=(63.91−58.59)/58.59）和 7.5%（=(58.81−54.73)/54.73）。

在 CASP11 的 48 个难预测的蛋白质上，对于短程、中程和长程接触本文提出的 TargetPCM 依然分别获得 57.78%，56.40% 和 38.92% 的最卓越的预测效果。因此表 4 和表 5 中的结果进一步表明了本文提出的 TargetPCM 的有效性。

2.4　案例研究

本文采用在 CASP11 测试集中的蛋白质（T0769-D1）来直观地表明本文所提出的 TargetPCM 相比起 MetaPSICOV[25] 具有更优越的预测性能。T0769-D1 包括 97 个残基（即 $L=97$）并且拥有 143 个真实的长程接触。图 2 根据长程接触，对于蛋白质 T0769-D1 分别绘制了 MetaPSICOV

和 TargetPCM 的真实的和预测的对比图。上三角的十字标记代表在 Top L 预测中的真实的长程接触，然而在下三角的十字标记代表在 Top L 预测中的预测的长程接触。

在 Top L 的预测中，MetaPSICOV 预测出了44个真实的长程接触，而 TargetPCM 正确预测出了56个真实的长程接触。在图2中，TargetPCM 比 MetaPSICOV 的预测性能更加优异。比如，在 MetaPSICOV 和 TargetPCM 的预测中存在两个假正样本数据簇(即图2的两个子图中的簇 A 和簇 B)。在两个簇中，MetaPSICOV 预测的长程接触的数量明显多于 TargetPCM，这表明在 MetaPSICOV 的长程接触预测中许多负样本被预测为正样本。

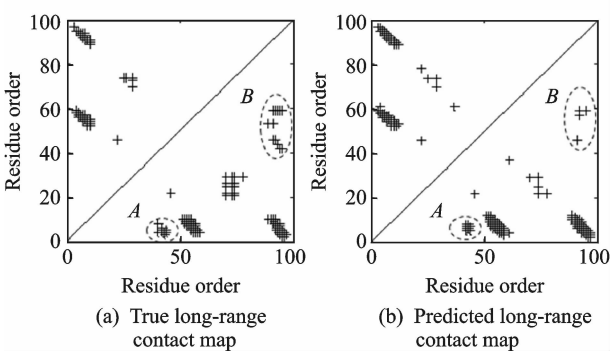

图2 MetaPSICOV 和 TargetPCM 对于蛋白质 T0769-D1 的真实的和预测的长程接触图

Fig. 2 True versus predicted long-range contact maps of MetaPSICOV and TargetPCM for protein T0769-D1

3 结束语

本文提出了一种基于 WNBC-PSO 和 ERT 的用于蛋白质残基接触图预测的集成预测器(TargetPCM)。TargetPCM 使用 WNBC-PSO 组合了 RFcon，PSICOV[13] 和 MetaPSICOV[25]，它们分别是基于机器学习、基于协同进化和基于集成的预测器，组合这些不同类型的预测器能够充分利用它们提供的互补信息。为了进一步提高预测性能，本文将 WNBC-PSO 的输出(即基于预测器的特征)和蛋白质基于序列的特征组合来形成更有判别力的特征，基于这个鲁棒的特征表示方法，训练得到最终的 ERT 预测模型(TargetPCM)。

为了验证 TargetPCM 的预测性能，本文将 TargetPCM 在 Test_98 和 CASP11 这两个独立的数据集上测试，发现对于短程、中程和长程接触，TargetPCM 在 Test_98 上的 Top $L/5$ 的预测精度比最好的集成预测器(NeBcon)分别提高了8.2%，16.1%和5.3%。在 CASP11 数据集上，对于短程、中程和长程接触，TargetPCM 的 Top $L/5$ 的预测精度比最好的基于协同进化的集成预测器(MetaPSICOV)分别提高了7.4%，9.1%和7.5%，这些比较表明了本文提出的 TargetPCM 对于蛋白质接触图预测的有效性。

尽管 TargetPCM 取得了良好的预测性能，但仍有进一步改进的空间。将来的工作将研究更有效的特征提取方法和机器学习方法(比如深度学习[40])来进一步提高蛋白质接触图的预测性能。

参考文献：

[1] CHENG J, BALDI P. A machine learning information retrieval approach to protein fold recognition[J]. Bioinformatics, 2006, 22(12): 1456-1463.

[2] OLMEA O, ROST B, VALENCIA A. Effective use of sequence correlation and conservation in fold recognition[J]. Journal of Molecular Biology, 1999, 293(5): 1221-1239.

[3] BONNEAU R, RUCZINSKI I, TSAI J, et al. Contact order and ab initio protein structure prediction [J]. Protein Science a Publication of the Protein Society, 2002, 11(8): 1937-1944.

[4] KLIGER Y, LEVY O, OREN A, et al. Peptides modulating conformational changes in secreted chaperones: From in silico design to preclinical proof of concept[J]. Proceedings of the National Academy of Sciences, 2009, 106(33): 13797-13801.

[5] PIERRE B, CHENG J. Improved residue contact prediction using support vector machines and a large feature set[J]. BMC Bioinformatics, 2007, 8(1): 1-9.

[6] YANG J, JIN Q Y, ZHANG B, et al. R2C: Improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter[J]. Bioinformatics, 2016, 32(16): 2435-2443.

[7] LI Y, FANG Y, FANG J. Predicting residue-residue contacts using random forest models[J]. Bioinformatics, 2011, 27(24): 3379-3384.

[8] GROMIHA M M. Influence of long-range contacts and surrounding residues on the transition state structures of proteins[J]. Analytical Biochemistry, 2011, 408(1): 32-36.

[9] MONASTYRSKYY B, DANDREA D, FIDELIS K, et al. Evaluation of residue-residue contact prediction in CASP10[J]. Proteins-Structure Function & Bioinformatics, 2014, 82(S2): 138-153.

[10] MONASTYRSKYY B, DANDREA D, FIDELIS K,

et al. New encouraging developments in contact prediction: Assessment of the CASP11 results[J]. Proteins-Structure Function & Bioinformatics, 2015, 84 (S1): 131-144.

[11] SKOLNICK J, KIHARA D, ZHANG Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm [J]. Proteins-Structure Function & Bioinformatics, 2004, 56(3): 502-518.

[12] MISURA K M, CHIVIAN D, ROHL C A, et al. Physically realistic homology models built with rosetta can be more accurate than their templates[J]. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(14): 5361-5366.

[13] JONES D T, BUCHAN D W, COZZETTO D, et al. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments [J]. Bioinformatics, 2012, 28 (2): 184-190.

[14] KAMISETTY H, OVCHINNIKOV S, BAKER D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110 (39): 15674-15679.

[15] KAJ N L, HOPF T A, KALAS M, et al. FreeContact: Fast and free software for protein contact prediction from residue co-evolution[J]. BMC Bioinformatics, 2014, 15(1): 85.

[16] SEEMAYER S, GRUBER M, S DING J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations[J]. Bioinformatics, 2014, 30(21): 3128-3130.

[17] CHENG J, BALDI P. Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms[J]. Bioinformatics, 2005, 21(S): 75-84.

[18] WU S, ZHANG Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction[J]. Bioinformatics, 2008, 24 (7): 924-931.

[19] DI L P, NAGATA K, BALDI P. Deep architectures for protein contact map prediction[J]. Bioinformatics, 2012, 28(19): 2449-2457.

[20] WANG Z, XU J. Predicting protein contact map using evolutionary and physical constraints by integer programming[J]. Bioinformatics, 2013, 29 (13): 266-273.

[21] MARKS D S, COLWELL L J, SHERIDAN R, et al. Protein 3D structure computed from evolutionary sequence variation [J]. PLOS One, 2011, 6 (12): e28766.

[22] TEGGE A N, WANG Z, EICKHOLT J, et al. NNcon: Improved protein contact map prediction using 2D-recursive neural networks[J]. Nucleic Acids Research, 2009, 37(Web Server issue): 515-518.

[23] YANG J, JANG R, ZHANG Y, et al. High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3D structure modeling[J]. Bioinformatics, 2013, 29(20): 2579-2587.

[24] SKWARK M J, RAIMONDI D, MICHEL M, et al. Improved contact predictions using the recognition of protein like contact patterns[J]. Plos Computational Biology, 2014, 10(11): e1003889.

[25] JONES D T, SINGH T, KOSCIOLEK T, et al. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins[J]. Bioinformatics, 2015, 31 (7): 999-1006.

[26] HE B, MORTUZA S M, WANG Y, et al. NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers [J]. Bioinformatics, 2017, 33(15): 2296-2306.

[27] WEBB G I, PAZZANI M J. Adjusted probability naive Bayesian induction[C]//Australian Joint Conference on Artificial Intelligence. [S. l.]: [s. n.], 1998: 285-295.

[28] ZHANG H, SHENG S. Learning weighted naive Bayes with accurate ranking[C]//IEEE International Conference on Data Mining. [S. l.]: IEEE, 2004: 567-570.

[29] 邓维斌，王国胤，洪智勇. 基于粗糙集的加权朴素贝叶斯邮件过滤方法[J]. 计算机科学，2011，38(2)：218-221.
DENG Weibin, WANG Guoyin, HONG Zhiyong. Weighted Naïve Bayes spam filtering method based on rough set[J]. Computer Science, 2011, 38(2): 218-221.

[30] EBERHART R C. A modified particle swarm optimizer[M]. Berlin, Heidelberg: Springer, 1998: 69-73.

[31] KENNEDY J, EBERHART R. Particle swarm optimization [C]//IEEE International Conference on Neural Networks, 1995 Proceedings. Perth, WA, Australia: IEEE, 2002: 1942-1948.

[32] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory[C]// International Symposium on MICRO Machine and Human Science. Nagoya, Japan: IEEE, 2002: 39-43.

[33] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63 (1): 3-42.

[34] 刘万里，刘三阳，王金艳. 不平衡支持向量机的调整

方法[J]. 计算机科学，2009，36(3)：148-149.
LIU Wanli，LIU Sanyang，WANG Jinyan. Adjusting method for imbalanced support vector machines[J]. Computer Science，2009，36(3)：148-149.

[35] 张洪强，刘光远，赖祥伟. 随机森林算法在肌电的重要特征选择中的应用[J]. 计算机科学，2013，40(1)：200-202.
ZHANG Hongqiang，LIU Guangyuan，LAI Xiangwei. Application of random forest algorithm in important feature selection from EMG signal[J]. Computer Science，2013，40(1)：200-202.

[36] 高尚，杨静宇. 一种新的基于粒子群算法的聚类方法[J]. 南京航空航天大学学报，2006，38(S1)：62-65.
GAO Shang，YANG Jingyu. New clustering method based on particle swarm algorithm[J]. Journal of Nanjing University of Aeronautics & Astronautics，2006，38(S1)：62-65.

[37] 白俊杰，王宁生，唐敦兵. 一种求解多目标柔性作业车间调度的改进粒子群算法[J]. 南京航空航天大学学报，2010，42(4)：447-453.
BAI Junjie，WANG Ningsheng，TANG Dunbing. Improved PSO algorithm for multi-objective optimization flexible job shop scheduling problems[J]. Journal of Nanjing University of Aeronautics & Astronautics，2010，42(4)：447-453.

[38] PEDREGOSA F，GRAMFORT A，MICHEL V，et al. Scikit-learn：Machine learning in python[J]. Journal of Machine Learning Research，2012，12(10)：2825-2830.

[39] LUND O，FRIMAND K，GORODKIN J，et al. Protein distance constraints predicted by neural networks and probability density functions[J]. Protein Engineering，1997，10(11)：1241-1248.

[40] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436-444.

（编辑：刘彦东）