Predicting protein condensate formation using machine learning

俞贞靖

摘要

无膜细胞器是通过液-液相分离形成的液体冷凝液。最近的进展表明，相分离通过调节包括转录和信号转导在内的基本细胞过程，对细胞内环境稳定至关重要。据报道，具有介导蛋白质相分离（PPS）能力的蛋白质数量不断增加。虽然已经开发了预测 PPS 的计算工具，但获得 PPS 概率的蛋白质组整体概览仍然具有挑战性。在此，该论文提出了一种相分离分析与预测（PSAP）机器学习分类器，该分类器仅基于已知 PPS 蛋白质训练集的氨基酸含量，可以确定给定蛋白质组中每个蛋白质的相分离可能性。通过与 PPS 数据库、现有预测因子和实验证据的比较，该论文证明了 PSAP 分类器的有效性和优势。该论文预计，PSAP 预测因子将为未来研究提供一个有用的工具，以识别健康和疾病中的相分离蛋白。

关键词：蛋白质相分离；机器学习

1 引言

细胞内存在大量缺乏物理边界的细胞器，通常被称为无膜细胞器或生物分子冷凝物。这些冷凝物具有类似液体的性质，可以与周围环境融合、交换组分，且在响应胞内和胞外信号时不稳定。液体冷凝物的形成和维持归因于一个称为相分离的过程。

近年来，通过相分离有潜力形成液体冷凝液的蛋白质数量迅速增加。蛋白质相分离通常是通过被称为内在无序区域的非结构域中折叠结构域或氨基酸基序之间的多价弱相互作用来介导的。目前的工作表明特定蛋白质的相分离潜力似乎编码在其氨基酸序列中。

鉴于蛋白质相分离 (PPS) 在细胞内稳态中的核心作用以及潜在的疾病治疗机会，评估任何给定生物体的整个蛋白质组的相分离能力将非常重要。虽然已经开发了预测 PPS 的计算工具，但获得 PPS 概率的蛋白质组整体概览仍然具有挑战性。因此本次复现的工作提出了一种相分离分析与预测（PSAP）机器学习分类器，该分类器仅基于已知 PPS 蛋白质训练集的氨基酸含量，可以确定给定蛋白质组中每个蛋白质的相分离可能性。通过与 PPS 数据库、现有预测因子和实验证据的比较，证明了 PSAP 分类器的有效性和优势。

2 相关工作

2.1 传统实验方法和数据库

在体外，研究者通常可以通过普通光学显微镜就能观察相分离现象，其特点为澄清的溶液会变得浑浊且能观察到类似油状的液滴。不仅如此，浊度测量与离心沉淀方法也被用于体外检测相分离。在体内通常还需要利用荧光漂白恢复技术（Fluorescence recovery after photobleaching，FRAP）来评估其流动性，从而鉴定蛋白质相分离的发生。对在生理条件下可发生相分离的蛋白质序列特征的研究表明，

在活细胞中，只有特定的蛋白质序列能够在适当的条件下进行相分离。因此，蛋白质的分子特性对于其潜在相分离行为的发生至关重要。

随着人们对相分离领域的广泛关注，建立了许多相分离相关数据库，表 1 汇总了这部分数据库。例如，Ning[1]等人收集了与相分离相关的蛋白并整合多物种同源蛋白，构建了一个名为 DrLLPS 的相分离相关蛋白数据库，其中包含了 164 个真核生物中的 7,993 个支架蛋白、72,300 个调控分子和 357,594 个客户蛋白，并对 8 种模式生物中与相分离相关的蛋白进行了详细的注释，如蛋白质翻译后修饰、无序区域、结构域、功能注释等。PhaSepDB 则是收集了 2,914 个相分离相关蛋白，并提供这些蛋白的发表来源、序列特征和免疫荧光图像等信息[2]。LLPSDB 数据库仅对 273 个实验验证的相分离蛋白进行了收集处理，详细地描述了每个蛋白的实验条件、功能注释等[3]。PhaSePro 是一个收集了实验验证的相分离驱动蛋白以及相关调控序列的数据库[4]。

表 1: 常见用于相分离预测的工具

数据库	链接
DrLLPS	http://llps.biocuckoo.cn/
PhaSepDB	http://db.phasep.pro/
LLPSDB	http://bio-comp.ucas.ac.cn/llpsdb
PhaSePro	https://phasepro.elte.hu

2.2 相分离传统预测工具

许多参与相分离的蛋白质已经被证明包含类似于朊病毒的区域或者含有内在无序区域（Intrinsically disordered regions，IDR）。因此，目前的研究中经常使用预测蛋白质朊病毒样区域的工具 PLAAC[5]、预测蛋白质序列无序区域的工具 IUPred[6]、PONDR-FIT[7]和 MobiDB[8]等来预测蛋白质相分离区域。此外，疏水性和带电残基被认为影响静电相互作用，这也被证实与相分离行为有关[9]。最近，Vernon 等对这些特征的预测性能进行了比较，并指出将多个特征结合起来可能有助于开发更准确的预测工具[10]。在表 2 中汇总了一些目前研究中常用的预测方法，到目前为止还没有一个工具通过整合多个序列特征来预测调控蛋白质相分离的关键氨基酸序列。

表 2: 常见用于相分离预测的工具

工具	描述	链接
MobiDB	预测低复杂区域	http://mobidb.bio.unipd.it/
IUPred	预测内在无序区	https://iupred2a.elte.hu/
PONDR-FIT	预测内在无序氨基酸	https://www.disprot.org/
D2P2	一种对蛋白质紊乱区域的预测	http://d2p2.pro/
DisMeta	蛋白质紊乱的元预测器	http://www-nmr.cabm.rutgers.edu/bioi nformatics/disorder/
Pi-Pi	预测粒子的驱动力	-
PLAAC	类朊病毒域预测器	http://plaac.wi.mit.edu
ZipperDB	预测纤维形成序列	https://services.mbi.ucia.edu/zipperdb/

2.3 机器学习算法的应用

近年来，随着计算机计算能力的不断提升，机器学习算法在自然语言处理、图像识别等领域都有着广泛的应用[11]。常见的机器学习算法包括隐马尔可夫模型[12]、支持向量机[13]、决策树[14]等，根据处理问题的不同，可以应用不同的机器学习算法。机器学习算法按照处理的问题不同主要分为三种类型，分别是分类、回归以及聚类。在分类问题中，通常需要设置给定样本标签，然后通过算法预测未知样本类别[15]。回归问题常被用于预测未来天气、房价等情况，用来推测某个具体数值[16]。而聚类问题则

是在不指定训练样本分类标签的情况下，通过算法将一些具有相似特征的数据进行聚集的过程[17]，例如利用 k-means 算法可以对不同来源的微生物进行聚类[18]。机器学习算法在多种生物学问题的处理中也有着广泛的应用[19]。在蛋白组学[20]的研究中，包括对蛋白质二级结构[21]和三级结构的预测[22]、蛋白质和蛋白质之间的相互作用[23]、蛋白质细胞亚定位[24]、蛋白质翻译后修饰位点等的预测[25]，如使用支持向量机模型对谷胱甘肽修饰位点进行预测的 GSHSite[26]，预测磷酸化位点的 GPS[27]和 iGPS[28]等；在基因组学[29]中，对转录因子起始位点的识别[30]、剪切位点的预测[31]以及功能结合位点的识别[32]等；不仅如此，机器学习也可参与到新药研发中[33]，例如 Iorio 等人使用方差分析、逻辑模型和机器学习算法（弹性回归网络和随机森林）来识别预测药物反应的分子特征及癌症药物间相互作用[34]。

3 本文方法

3.1 本文方法概述

使用已知的人类相分离蛋白列表，将其与其他人类蛋白质组进行比较，生成一系列氨基酸相关特征，使用这些特征开发了一个随机森林分类器（PSAP），并使用 10 折交叉验证来评估其性能。使用 PSAP 模型预测蛋白质组中每个蛋白质的相分离概率，并将其与 Pscore 进行比较。

3.2 特征提取模块

对从 Uniprot 数据库中获得的 90 个 pps 蛋白质的序列和其余蛋白质组进行以下特征的提取：使用 Biopython 包中的 ProteinAnalysis 模块来计算 20 种氨基酸的百分比、长度、等电点、分子量、肉汤（gravy）、芳香性以及 alpha 螺旋、beta 转角和 beta 薄片的分数；氨基酸的可变基团属性:Asx: D, N, Glx: E, Q, Xle: I, L，正电荷:K, R, H，负电荷:D, E，芳香:F, W, Y, H，脂肪族:V, I, L, M，小:P, G, A, S，亲水:S, T, H, N, Q, E, D, K, R，疏水:V, I, L, F, W, Y, M；通过滑动窗口大小为 20 中氨基酸（≤7）的数量来确定低复杂性得分（LCS），使用 Seaborn catplot 计算并可视化这些低复杂度区域中的每种氨基酸的数量。使用 IUPred2A 全局算法计算蛋白质序列每个位置的内在无序得分，得分高于 0.5、0.6、0.7、0.8 和 0,9 的氨基酸的比例；计算每种蛋白质的 Kyte 和 Doolittle 分数得到疏水值，疏水值低于 1.5、2.0 或 2.5 的氨基酸的分数和总数；以上共提取了 95 个特征。

3.3 模型构建模块

数据通过使用来自 Scikit learn 的 StandardScaler 进行缩放归一化。在从这些序列中提取的特征中，去除 Pearson 相关性低于 0.95 的特征。随机森林的超参数选择如下：RandomForestClassifier: n_estimators = 1200(树的数量)，class_weight = ' balanced ' 和 criterion = ' entropy '。从这个随机森林模型中提取特征重要性，以选择模型认为对其预测很重要的特征，主要包括 L 和 C 的百分比、低复杂度分数（LCS）和 IDR 的百分比。

3.4 性能评估模块

随机森林模型用 sklearn AUC metric 计算了接收工作特征 (ROC) 的曲线下面积 (AUC)。此外，使用 klearn 的 precision_score 和 recall_score 计算精度和召回率，使用 precision_recall_curve 计算精度召回曲线 (PRC) 的 AUC。所有的性能评估都基于 10 倍交叉验证。为了进一步评估 PSAP 的性能，将 PSAP 与 Pscore 进行了深入比较，Pscore 是人类蛋白质组中 PPS 蛋白质的预测工具，是性能最好的第一代相分离预测因子。首先对蛋白质组的 PSAP 和 Pscore 预测分数进行排名，在数据库中注释为相分

离的蛋白质在 PSAP 预测中排名较高；其次对支架蛋白质、调节器蛋白质、以及这些类别中不存在的蛋白质（“其他”）进行相分离预测，支架和客户在 PSAP 预测中排名较高，而“调节”或“其他”蛋白质没有增加。最后，使用 ROC 和 PRC 分析评估了 PSAP 和 Pscore 模型，与 Pscore 相比，PSAP 分类器在对 90 个高置信度相分离蛋白集进行分类方面总体表现更好。

4 复现细节

4.1 与已有开源代码对比

该论文的代码是开源的，但不提供数据集以及 Pscore 的实现，因此从已有的数据库构建自己的数据集，其中包含 90 个经过实验验证能发生相分离的蛋白，总共 20380 个蛋白质序列。此外实现了 Pscore 对这 20380 个蛋白质序列是否发生相分离的预测。

4.2 实验环境搭建

创建 python=3.6 的虚拟环境，下载 iupred2a 包，在 vscode 编译器上提取蛋白质序列的特征，在 jupyter notebook 上构建随机森林模型并对结果进行可视化。

4.3 界面分析与使用说明

使用 fasta2feature.py 文件对蛋白质序列进行特征提取；使用 add_labels.ipynb 文件对蛋白质序列添加标签；使用 rf.ipynb 文件构建随机森林模型对蛋白质序列进行相分离预测。

4.4 创新点

无

5 实验结果分析

特征提取模块：

```
[Running] python -u "d:\jupyter_file\iupred2a\test.py"
self.fasta2df(dbfasta)
20380
0.36s self.fasta2df(dbfasta)
self.amino_acid_analysis()
```

图 1: 氨基酸

```
13.66s self.amino_acid_analysis()
self.idr_iupred()
```

图 2: 无序区

```
317.88s self.idr_iupred()
self.hydrophobic()
13.51s self.hydrophobic()
self.add_iupred_features()
```

图 3: 疏水值

```
14.31s self.add_iupred_features()
self.add_hydrophobic_features()
```

图 4: 疏水值特征

```
144.03s self.add_hydrophobic_features()
self.add_biochemical_combinations()
0.59s self.add_biochemical_combinations()
self.add_lowcomplexity_features()
```

图 5: 生物化学组合

```
108.48s self.add_lowcomplexity_features()
Generated file: total_llps_f2f_5-12-2022.pkl
```

图 6: 低复杂区域特征

性能评估模块：

100% 10/10 [01:52<00:00, 11.04s/it]

```
Fold 1 ROC AUC:  0.9702
Fold 2 ROC AUC:  0.8108
Fold 3 ROC AUC:  0.9838
Fold 4 ROC AUC:  0.7008
Fold 5 ROC AUC:  0.7441
Fold 6 ROC AUC:  0.9670
Fold 7 ROC AUC:  0.8792
Fold 8 ROC AUC:  0.9282
Fold 9 ROC AUC:  0.9044
Fold 10 ROC AUC:  0.7796
```

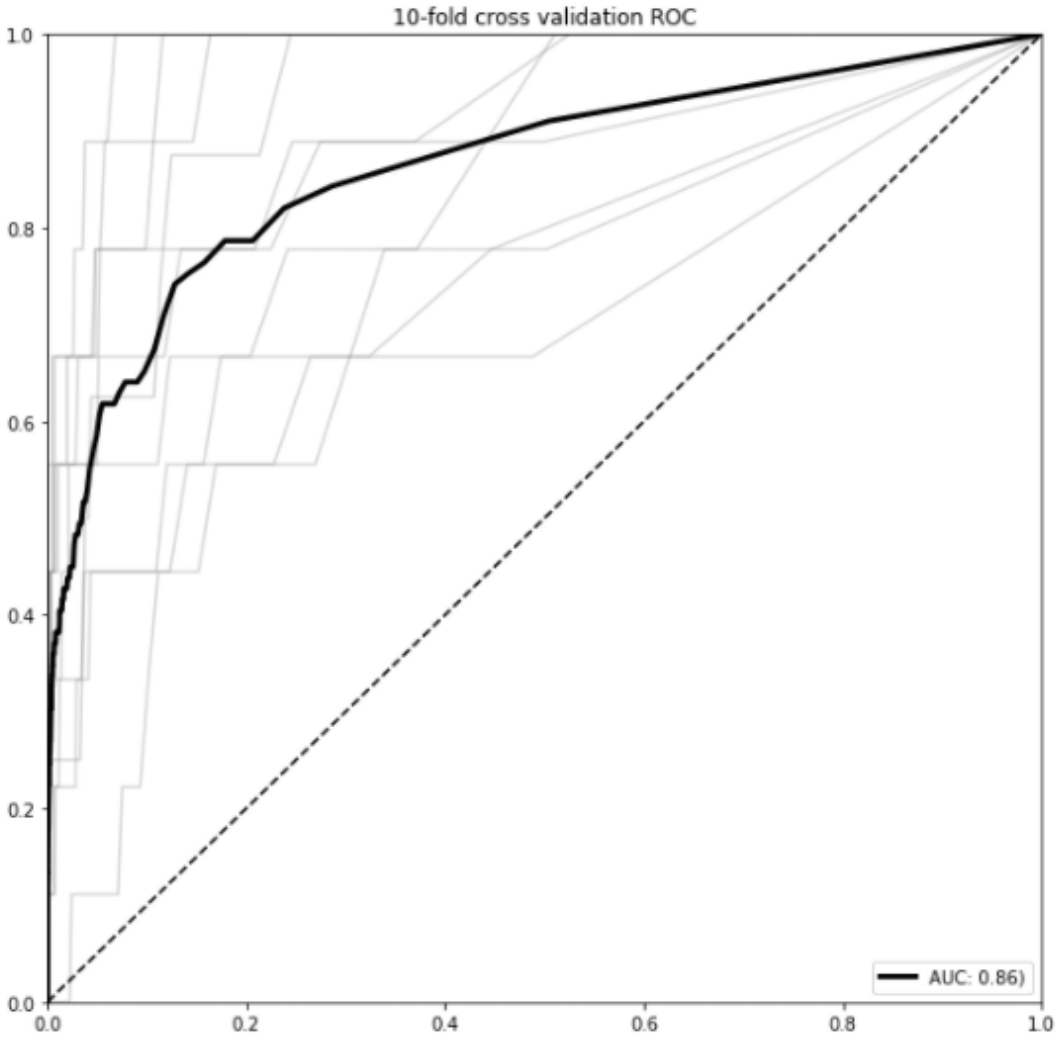

图 7: ROC 曲线

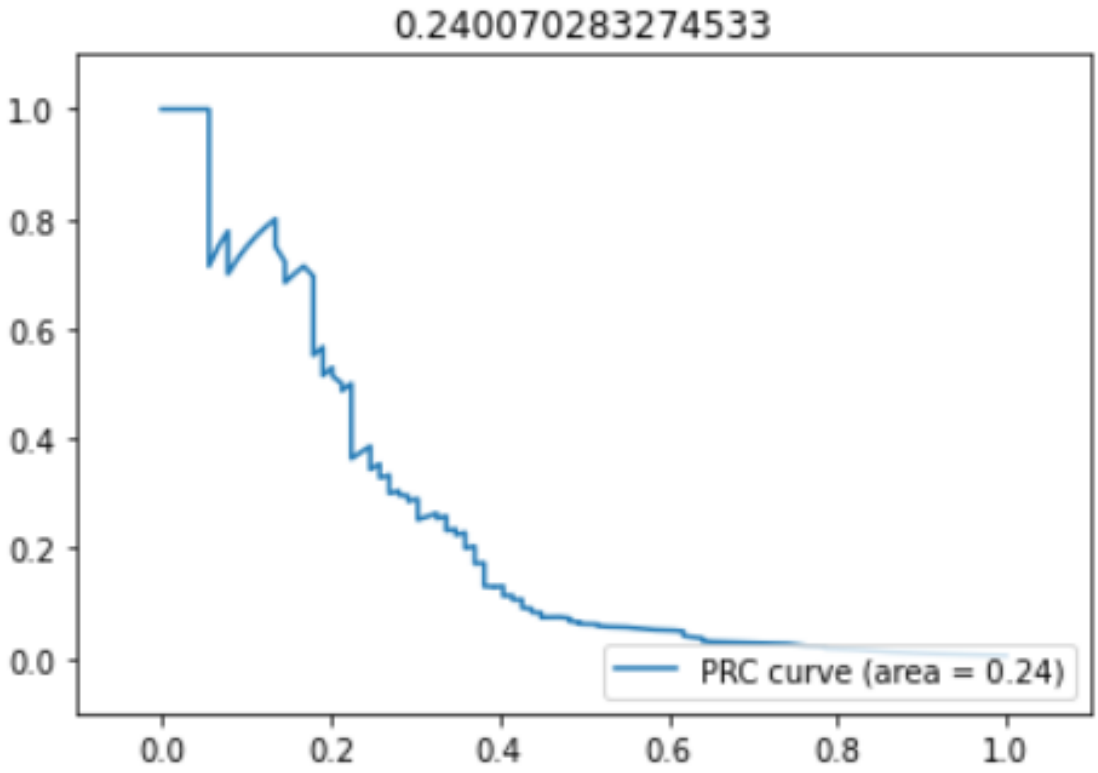

图 8: PRC 曲线

获取蛋白质序列的 pscore 得分：

```
PS D:\paper\elife-31486-code2-v2\SourceCodeS2> python elife_phase_separation_predictor.py 20380.fasta -output pscore2.out
Loading Database...
Working on sequence:        1 of 20380
Working on sequence:        2 of 20380
Working on sequence:        3 of 20380
Working on sequence:        4 of 20380
Working on sequence:        5 of 20380
Working on sequence:        6 of 20380
Working on sequence:        7 of 20380
Working on sequence:        8 of 20380
Working on sequence:        9 of 20380
Working on sequence:       10 of 20380
```

图 9: Pscore 得分 1

```
Working on sequence:    20366 of 20380
Working on sequence:    20367 of 20380
Working on sequence:    20368 of 20380
Working on sequence:    20369 of 20380
Working on sequence:    20370 of 20380
Working on sequence:    20371 of 20380
Working on sequence:    20372 of 20380
Working on sequence:    20373 of 20380
Working on sequence:    20374 of 20380
Working on sequence:    20375 of 20380
Working on sequence:    20376 of 20380
Working on sequence:    20377 of 20380
Working on sequence:    20378 of 20380
Working on sequence:    20379 of 20380
Working on sequence:    20380 of 20380
```

图 10: Pscore 得分 2

PSAP 与 Pscore 性能对比：

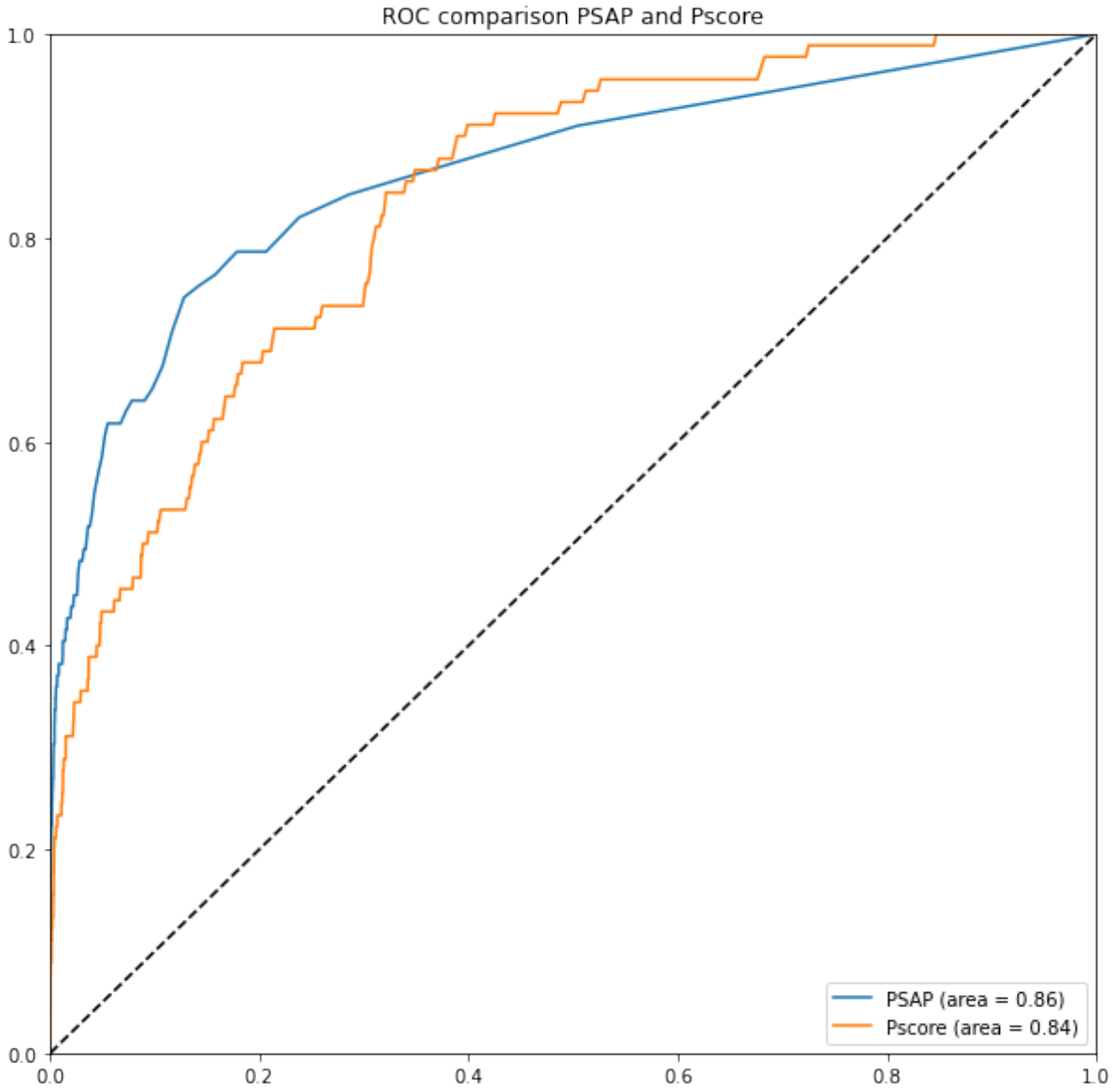
ROC comparison PSAP and Pscore
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PSAP (area = 0.86)
Pscore (area = 0.84)

图 11: ROC 曲线

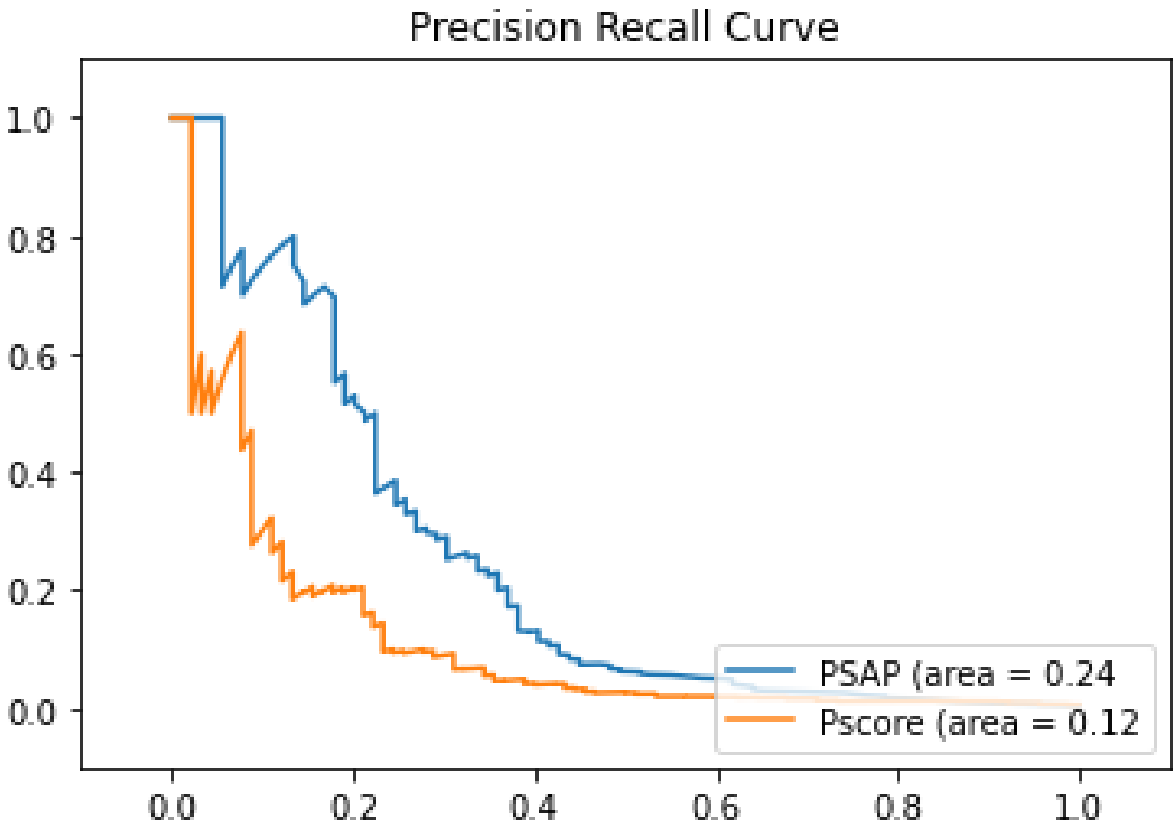

图 12: PRC 曲线

6 总结与展望

本部分对整个文档的内容进行归纳并分析目前实现过程中的不足以及未来可进一步进行研究的方向。PPS 最近作为一个对细胞内环境稳定至关重要的过程而受到关注。因此，人们正在努力了解 PPS 的分子原理。这导致了一些蛋白质特征的识别，例如 RG 延伸，这是 PPS 的重要决定因素。之前的研究旨在利用这些蛋白质的一个或几个特征来区分 PPS 蛋白质和其他蛋白质组。然而，这些所谓的第一代相分离预测仪并没有综合考虑与相分离相关的多种特征，这突出表明需要改进预测仪。为了改进这些现有方法，该论文从实验验证的 PPS 蛋白的 aa 序列中定义了特征，而无需事先了解这些特征的组成，并开发了 PSAP 预测器。与迄今为止表现最好的第一代预测因子 Pscore 相比，PSAP 预测因子将更高的预测概率归因于 PPS 数据库中的蛋白质。因此，与现有分类器相比，PSAP 是一种非常有价值且经过改进的替代方法，并且可能是第二代预测工具中的第一个。

虽然这项研究的重点是人类蛋白质组，但 PSAP 可以应用于每个有 FASTA 蛋白质组文件的生物体，以及一组已知的经实验验证的 PPS 蛋白质，这些蛋白质已在最近的数据库中为各种生物体保存。因此，PSAP 将为相分离领域寻找新的相分离蛋白提供一个有用的工具。此外，PSAP 或许可用于预测疾病相关突变的影响，如癌症中发生的突变，或 IDR 中经常发生的突变，从而导致异常的二肽，如在所谓的双亮氨酸病中的亮氨酸二肽，这是未来可进一步研究的方向。

参考文献

[1] Ning W, Guo Y, Lin S, et al. DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes [J]. Nucleic Acids Research, 2019, 48(D1): D288-D295.

[2]You K, Huang Q, Yu C, et al. PhaSepDB: a database of liquid–liquid phase separation related proteins [J]. Nucleic Acids Research, 2019, 48(D1): D354-D359.

[3] Li Q, Peng X, Li Y, et al. LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro [J]. Nucleic Acids Research, 2019, 48(D1): D320-D327.

[4] Mészáros B, Erdős G, Szabó B, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation [J]. Nucleic Acids Research, 2020, 48(D1): D360-D367.

[5] Lancaster A K, Nutter-Upham A, Lindquist S, et al. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition [J]. Bioinformatics, 2014, 30(17): 2501-2502.

[6] Dosztanyi Z, Csizmok V, Tompa P, et al. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content [J]. Bioinformatics, 2005, 21(16): 3433-3434.

[7] Xue B, Dunbrack R L, Williams R W, et al. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids [J]. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2010, 1804(4): 996-1010.

[8] Piovesan D, Tabaro F, Paladin L, et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins [J]. Nucleic Acids Research, 2017, 46(D1): D471-D476.

[9] Aumiller W M, Keating C D. Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles [J]. Nature Chemistry, 2015, 8(2): 129-137.

[10] Vernon R M, Forman-Kay J D. First-generation predictors of biological protein phase separation [J]. Current Opinion in Structural Biology, 2019, 58: 88-96.

[11] 刘太岗. 机器学习方法在生物信息学中的应用 [D]; 大连理工大学, 2010.

[12]Rabiner L R. A tutorial on hidden markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[13] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[14] Quinlan J R. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81-106.

[15] 崔霞霞. 基于机器学习的分类问题研究 [D]; 中北大学, 2018.

[16] 杨浩浩. 几种机器学习算法及其集成模型在回归问题中的应用与比较 [D]; 兰州大学, 2018.

[17] Hamerly G, Elkan C J a I N I P S. Learning the k in k-means [J]. Advances in neural information processing systems, 2004, 16: 281-288.

[18] Goodswen S J, Barratt J L N, Kennedy P J, et al. Machine learning and applications in microbiology [J]. FEMS Microbiology Reviews, 2021.

[19] Larrañaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics [J]. Briefings in Bioinformatics, 2006, 7(1): 86-112.

[20] Lin H, Chen W, Anandakrishnan R, et al. Application of machine learning method in genomics and proteomics [J]. ScientificWorldJournal, 2015, 2015: 914780.

[21] King R D, Sternberg M J. Machine learning approach for the prediction of protein secondary structure [J]. Journal of Molecular Biology, 1990, 216(2): 441-457.

[22] Hou J, Wu T, Cao R, et al. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13 [J]. Proteins, 2019, 87(12): 1165-1178.

[23] Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure [J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(21): 12105-12110.

[24] Qi Y, Klein-Seetharaman J, Bar-Joseph Z. Random forest similarity for protein-

protein interaction prediction from multiple sources [J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2005: 531-542.

[25] Horton P, Park K J, Obayashi T, et al. WoLF PSORT: protein localization predictor [J]. Nucleic Acids Research, 2007, 35(Web Server issue): W585-587.

[26] Mohabatkar H, Rabiei P, Alamdaran M. New achievements in bioinformatics prediction of post translational modification of proteins [J]. Current Topics in Medicinal Chemistry, 2017, 17(21): 2381-2392.

[27] Chen Y-J, Lu C-T, Huang K-Y, et al. GSHSite: exploiting an iteratively statistical method to identify S-glutathionylation sites with substrate specificity [J]. PLoS One, 2015, 10(4): e0118752.

[28] Xue Y, Zhou F, Zhu M, et al. GPS: a comprehensive www server for phosphorylation sites prediction [J]. Nucleic Acids Research, 2005, 33(suppl_2): W184-W187.

[29] Song C, Ye M, Liu Z, et al. Systematic analysis of protein phosphorylation networks from phosphoproteomic data [J]. Molecular & Cellular Proteomics, 2012, 11(10): 1070-1083.

[30] Libbrecht M W, Noble W S. Machine learning applications in genetics and genomics [J].Nature Reviews Genetics, 2015, 16(6): 321-332.

[31] Ohler U, Liao G C, Niemann H, et al. Computational analysis of core promoters in the Drosophila genome [J]. Genome Biology, 2002, 3(12): Research0087.

[32] Jaganathan K, Kyriazopoulou Panagiotopoulou S, Mcrae J F, et al. Predicting splicing from primary sequence with deep learning [J]. Cell, 2019, 176(3): 535-548.e524.

[33] Wang K, Jian Y, Wang H, et al. RBind: computational network method to predict RNA binding sites [J]. Bioinformatics, 2018, 34(18): 3131-3136.

[34] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development [J]. Nature Reviews Drug Discovery, 2019, 18(6): 463-477.