DOI:10.12113/202202013
网络首发地址:http://kns.cnki.net/kcms/detail/23.1513.Q.20220616.0955.004.html

基于支持向量机识别亚高尔基体蛋白质的定位

闫　婷,李凤敏*
(内蒙古农业大学 理学院,呼和浩特 010018)

摘　要:研究表明,许多神经退行性疾病都与蛋白质在高尔基体中的定位有关,因此,正确识别亚高尔基体蛋白质对相关疾病药物的研制有一定帮助,本文建立了两类亚高尔基体蛋白质数据集,提取了氨基酸组分信息、联合三联体信息、平均化学位移、基因本体注释信息等特征信息,利用支持向量机算法进行预测,基于 5-折交叉检验下总体预测成功率为 87.43%。

关键词:亚高尔基体;蛋白质;氨基酸组分;基因本体;支持向量机
中图分类号:Q61　　**文献标志码**:A　　**文章编号**:1672-5565(2023)01-045-06

Identification of sub-Golgi proteins localization based on support vector machine

YAN Ting, LI Fengmin*
(*College of Science*, *Inner Mongolia Agricultural University*, *Hohhot* 010018, *China*)

Abstract: Many neurodegenerative diseases are associated with the location of proteins in the Golgi apparatus. Therefore, the correct identification of sub-Golgi proteins is helpful for the development of drugs for related diseases. In this study, two types of sub-Golgi protein datasets were established. On the basis of the amino acid composition information, the conjoint triad feature information, the auto-covariance average chemical shift, and the gene ontology information, the localization of sub-Golgi protein was predicted by using the algorithm of support vector machine. The overall prediction accuracy was 87.43% in the 5-fold cross-validation.

Keywords: sub-Golgi; Protein; Amino acid composition; Gene ontology; Support vector machine

研究发现对基因进行修饰改变,可能会直接引发各类疾病,作为基因修饰功能实施者的蛋白质,在基因改变致病的过程中起着重要作用。所以了解蛋白质的功能十分重要[1]。对于生物体内存在的蛋白质,如果知道它在细胞中的位置,就能了解该蛋白质的生物学功能[2]。研究表明,帕金森病[3]和阿尔茨海默病[4]等疾病都与蛋白质的亚高尔基体位置相关,正确识别高尔基体驻留蛋白的类型,对于了解高尔基体蛋白质的功能有着非常重要的作用。亚细胞分离等实验识别亚高尔基体蛋白质位置的方法既费时又费钱,而且实验过程中还会出现很多问题。利用蛋白质序列及结构信息预测其亚细胞位置,成本低而且不耗费时间。

高尔基体是由许多扁平的囊泡构成的以分泌为主要功能的细胞器,是真核细胞中内膜系统的组成之一,它不仅存在于动植物细胞中,而且也存在于原生动物和真菌细胞内。高尔基体由三个扁平的膜囊组成,包括:顺面膜囊、中间膜囊和反面膜囊,结构如图 1 所示。

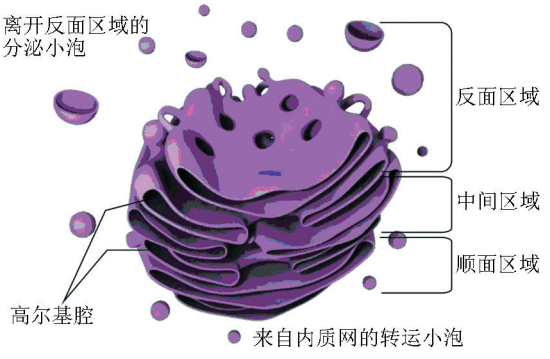

图 1　亚高尔基体的结构

Fig.1　Structure of sub-Golgi apparatus

收稿日期:2022-02-23;**修回日期**:2022-03-28;**网络首发日期**:2022-06-16.
基金项目:内蒙古自治区自然科学基金项目(No.2019MS03015).
作者简介:闫婷,女,硕士研究生,研究方向:理论生物物理. E-mail:276047868@qq.com.
*通信作者:李凤敏,女,教授,博士,研究方向:理论生物物理. E-mail:lfmbs@126.com.

高尔基体[5]在蛋白质糖基化、膜上的转化作用以及溶酶体的形成和植物细胞壁的形成等方面都发挥着重要作用。对亚高尔基体蛋白质的定位预测有助于理解高尔基体内发生的各种生物过程的机制,并了解亚高尔基体蛋白质的功能,许多疾病发生的致病机理都与亚高尔基体的位置相关,因此,亚高尔基体蛋白质的定位预测也可以为与高尔基体相关的疾病设计药物提供一定的帮助[6]。2008 年,van Dijk 等[7]对 II 型跨膜蛋白的单通道跨膜结构域(TMD)序列进行了预测,证明了亚高尔基体蛋白质的位置与聚糖生物合成中不同步骤的顺序之间有明显的关系。2011 年 Ding Hui 等[8]建立了两类亚高尔基体蛋白质的数据集并将离散增量(ID)和信息熵(IH)与改进的马氏判别式相结合对高尔基体蛋白质类型进行预测,预测成功率为 74.7%。2013 年,Ding Hui 等[9]提取了间隔二肽组分(g-gap dipeptide)信息来识别高尔基体蛋白质的类型。2015 年,Jiao Yasen 等[10]。基于 Ding Hui 的数据集,对 cis-Golgi 蛋白和 trans-Golgi 蛋白进行了区分,预测成功率为 86.9%。2016 年,Yang Runtao 等[11]基于 SwissProt 数据库建立了新的亚高尔基体蛋白质数据集,他们基于共同空间模式(CSP)的概念,开发了一种新的特征提取技术从蛋白质序列中提取进化信息,预测成功率为 86.4%。近几年,一些研究者对亚高尔基体蛋白质进行了深入研究,对亚高尔基体蛋白质的 cis-Golgi 蛋白和 trans-Golgi 蛋白进行预测,得到了较高的成功率[12-15]。2019 年,Zhao Wei 等[16]建立了三类亚高尔基体蛋白质的数据集,将伪氨基酸组分信息与功能域富集得分(FunDES)结合,预测成功率为 78.4%。2020 年,Cui Qingyu 等[17]。在特征提取中没有使用完整的蛋白质序列,提出了 529 种切割的蛋白质序列,训练集和测试集都根据这 529 种切割类型进行切割,预测成功率为 78.13%。

本文基于最新版本 SwissProt 数据库建立了包含顺面膜囊(cis-Golgi)和反面膜囊(trans-Golgi)两类亚高尔基体蛋白质的数据集,通过输入多种特征信息进行预测实验,最后选取 6 种预测结果较好的特征信息,6 种特征信息中基因本体(GO)注释信息的预测结果最好。在此基础上对 6 种特征信息进行融合,利用支持向量机算法对两类亚高尔基体蛋白质定位进行预测,融合特征最高预测成功率为 87.43%,特征信息的融合对预测结果有一些提升作用,本文的预测方法对反面膜囊(trans-Golgi)蛋白质的识别效果较好。

1 材料与方法

1.1 数据集的构建

在过去的几十年中,通过计算方法对蛋白质进行定位预测成为了生物信息学研究的重点,本文基于 Swiss-Prot 数据库,建立了两类亚高尔基体蛋白质的数据集,建立该数据集有以下几个步骤:

1)在 Swiss-Prot 数据库中搜索关键词“Golgi apparatus”,选择经过实验验证的、亚细胞位置在“cis-Golgi”和“trans-Golgi”的蛋白质。

2)去除 Fragment 序列。

3)去掉具有模糊蛋白质注释的序列,例如含有“By similarity”和“Probably”等注释的序列。

4)删除重复位置的蛋白质序列。

经过上述步骤得到 977 条亚高尔基体蛋白质,使用 CD-HIT 在线工具去除序列相似性,由于两类亚高尔基体蛋白质的数目差距较大,当阈值设为 25%时,顺面膜囊(cis-Golgi)蛋白质的数目较少,会导致两类亚高尔基体蛋白质数据集更加不平衡,所以我们将阈值设为 40%,得到两类亚高尔基体蛋白质的数据集,见表 1。

表 1 亚高尔基体蛋白质数据集中序列数目

Table 1 Number of sequences in sub-Golgi protein dataset

Sub-Golgi localization	Number of sequence
cis-Golgi proteins	74
trans-Golgi proteins	284
Total	358

2016 年Yang 等[11]基于 Swiss-Prot 数据库建立的两类亚高尔基体蛋白质数据集,该数据库被研究者广泛应用,这个数据集共 304 条蛋白质序列,其中包括 87 条顺面膜囊(cis-Golgi)蛋白质序列和 217 条反面膜囊(trans-Golgi)蛋白质序列。但是随着测序技术的发展,Swiss-Prot 数据库不断更新,顺面膜囊(cis-Golgi)蛋白质序列和反面膜囊(trans-Golgi)蛋白质序列数目不断变化,本文基于最新版本 Swiss-Prot 数据库建立了两类亚高尔基体蛋白质定位数据集,新建的数据集与 Yang 等[11]构建的数据集相比亚高尔基体蛋白质序列数目增加 54 条。

1.2 特征参数的选取

1.2.1 氨基酸组分信息

单肽组分信息(AAC)是对每一条蛋白质中 20 种氨基酸出现的概率进行统计得到的[18]。单肽组分信息可表示如下:

$$P = [v_1, v_2, v_3, \cdots v_i \cdots, v_{20}] \quad (1)$$

$$v_i = \frac{n_i}{L} \quad (2)$$

其中,L 表示蛋白质序列的长度,n_i 表示蛋白质中第 i 个氨基酸出现的个数。

二肽组分信息[19]。(DC)是通过计算每两个紧邻氨基酸出现的概率得到的。对于给定的蛋白质,二肽组分信息可表示如下:

$$F = [f_1, f_2, \cdots f_i, \cdots f_{400}] \quad (3)$$

$$f_i = \frac{n_i}{L - 1} \quad (4)$$

其中,L 是蛋白质序列的长度,n_i 表示蛋白质中第 i 个二肽出现的个数。

1.2.2　两亲性伪氨基酸组分信息

两亲性伪氨基酸组分(APAAC)是由 Chou 提出的[20],可以有效地识别蛋白质,并已广泛应用于不同蛋白质的序列分析。与传统 AAC 不同的是,APAAC 利用蛋白质中氨基酸的疏水性和亲水性,将蛋白质序列顺序与传统 AAC 结合起来。APAAC 可以表示如下:

$$P_{APAAC} = [P_1, P_2, \cdots, P_{20}, P_{20+1}, P_{20+2} \cdots P_{20+\lambda}, \cdots P_{20+2\lambda}] \quad (5)$$

其中,前 20 个向量代表的是传统的 AAC,λ 表示相关系数。

1.2.3　联合三联体特征

联合三联体特征(CTF)是 Shen 等人[21]基于预测蛋白质相互作用所提出的特征信息,他们考虑了一个氨基酸及其紧邻氨基酸的性质,20 种氨基酸被分为 7 类,见表 2。CTF 具体可表示为:

$$P_{CTF} = [f_1, f_2, f_3, \cdots f_i \cdots, f_{343}] \quad (6)$$

$$f_i = \frac{k_i}{L - 2} \quad (7)$$

其中 k_i 是每种联合三联体在蛋白质中出现的个数,L 是蛋白质序列的长度。

表 2　氨基酸分类

Table 2　Classification of amino acids

1	2	3	4	5	6	7
A,G,V	F,I,L,P	M,S,T,Y	H,N,Q,W	K,R	D,E	C

1.2.4　平均化学位移

本文通过将蛋白质序列提交到 PSIPRED 网站获得亚高尔基体蛋白质的二级结构,然后将蛋白质序列和其相对应的二级结构提交到 Fan 等人构建的平均化学位移[22]服务网站 acACS(http://202.207.14.87:8032/ bioinformation/acACS/index. asp)中,得到化学位移的结果,表示如下:

$$ACS_i^k(j) = \frac{1}{N} \sum \omega_i^k(j) \quad (8)$$

其中 i 表示 4 种骨架原子(^{15}N、^{13}C、^{1}H、$^{1}H_N$),k 表示蛋白质二级结构的类别(H、E、C),j 表示 20 种氨基酸,N 表示蛋白质序列中氨基酸的个数。

对于给定蛋白质 P,P 可表示如下:

$$P = [C_1^i, C_2^i, \cdots C_L^i](i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N) \quad (9)$$

该化学位移的自相关协方差可表示如下:

$$\psi_\lambda^i = \frac{1}{L - \lambda} \sum_{l}^{1-\lambda} [C_1^i - C_{1+\lambda}^i]^2 (i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N; 0 < \lambda < L) \quad (10)$$

则 PacACS 可表示如下:

$$P_{acACS} = [\psi_0^i, \psi_1^i, \psi_2^i, \cdots \psi_\lambda^i](i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N; 0 < \lambda < L) \quad (11)$$

其中 λ 为相关长度,L 为蛋白质序列的长度。对于不同蛋白质,要选择不同的骨架原子组合及合适的 λ 值,得到预测结果最好的特征参数。

1.2.5　基因本体注释信息

基因本体(GO)注释信息包括分子功能(F)、生物过程 P)及细胞组分(C)三部分[23],由于细胞组分(C)包含亚高尔基体蛋白质亚细胞位置的注释信息,为了预测结果的客观性,本文使用的 GO 注释信息去掉了细胞组分(C),本文提取的 GO 特征信息只包含生物过程(P)和生物功能(F)两个注释信息。GO 注释信息可以描述基因和基因产物的功能,许多研究中通过检测蛋白质或者遗传基因的 GO 号来提取基因本体注释信息,如果蛋白质的 GO 号未知,则使用同源序列的 GO 号,同源序列可由 BLAST 软件检测得到。

本文通过 AC 号在 Swiss-Prot 数据库中找到每条蛋白质对应的生物过程(P)和分子功能(F)的注释信息,没有 GO 注释信息的利用 BLAST 软件查找该序列的同源序列,同源阈值设为 60%,下载所有亚高尔基体蛋白质序列的 GO 注释信息后,整理、去重复后得到 1646 个 GO 注释信息,将对应位置有 GO 注释信息的设为 1,对应位置没有 GO 注释信息的设为 0,得到特征向量,表示如下:

$$P_{GO} = [f_1, f_2 \cdots f_n \cdots f_{1646}] \quad (12)$$

其中, 是指在位置 n,该蛋白质序列有无 GO 注释信息,它的定义如下:

$$f_n = \begin{cases} 1 & \text{在该位置存在 GO 信息} \\ 0 & \text{在该位置不存在 GO 信息} \end{cases} \quad (13)$$

1.3　预测算法

支持向量机(SVM)算法最早由 Vapnik 等[24]提出,已成功用于蛋白质结构和功能的预测。该算法的核心理念是将数据从低维向量映射到高维向量,

在高维空间中使正集和负集之间的距离最大化。本文使用 LibSVM 软件包对亚高尔基体蛋白质进行预测。

1.4 评价指标

本文采用 5-折交叉检验来评估预测性能。选用预测成功率(Accuracy, *Acc*)、敏感性(Sensitivity, *Sn*)、特异性(Specificity, *Sp*)和 Matthew 相关系数(Matthew's correlation coefficient, *MCC*)作为评价指标对算法进行评价[25]。评价指标的定义如下:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$Sn = \frac{TP}{TP + FN} \quad (15)$$

$$Sp = \frac{TN}{TN + FP} \quad (16)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + TN) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad (17)$$

其中,TP 表示真阳性,TN 表示真阴性,FP 表示假阳性,FN 表示假阴性。

2 结果与讨论

2.1 特征向量参数选取

两亲性伪氨基酸组分信息(APAAC)提取时,要选取不同 λ 值进行预测,当 λ = 30 时,预测结果最好;平均化学位移(acACS)提取时,要选取不同 λ 值和不同骨架原子组合进行预测,当 λ = 30,骨架原子组合为 $^{15}N+^{13}C_{\alpha}+^{1}H_{N}$ 时,对亚高尔基体蛋白质的预测结果最好。

2.2 预测结果

2.2.1 单特征预测结果

本文采用支持向量机算法,利用提取的 6 类不同特征信息,对两类亚高尔基体蛋白质进行预测,单特征的预测结果见表 3。

表 3 不同特征参数的预测结果

Table 3 Prediction results of different feature parameters

特征	Sn/%	Sp/%	Acc/%	MCC
AAC	97.17	13.33	80.01	0.196
DC	96.82	21.33	81.28	0.290
APAAC	100.	4.000	80.17	0.179
CTF	97.53	6.670	79.61	0.095
acACS	99.29	16.00	81.84	0.321
GO	95.40	54.67	87.15	0.569

由表 3 可以看出,单个特征信息预测成功率 *Acc*(除 CTF 外)都在 80%以上,所有特征信息的敏感性 *Sn* 都较高,最高可达到 100%,说明针对本文建立的两类亚高尔基体蛋白质数据集,该预测方法对反面膜囊(Trans-Golgi)蛋白质的预测结果较好,可以有效地识别反面膜囊(Trans-Golgi)蛋白质,而特异性 *Sp* 的结果都偏低,说明本文的方法对顺面膜囊(Cis-Golgi)蛋白质的预测结果较差。

为了直观的比较单个参数的预测成功率,以特征参数为横坐标,预测成功率 Acc 为纵坐标作图,图 2 表示不同特征参数的预测成功率 Acc,由图 2 可知,基因本体 GO 注释信息的预测成功率 Acc 最高,为 87.15%。

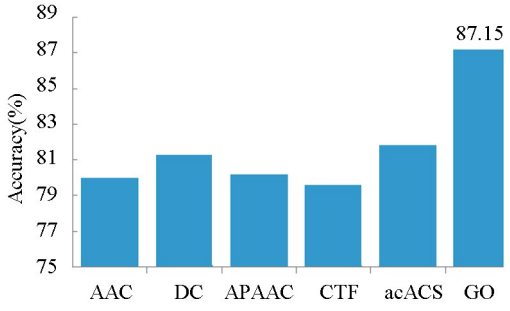

图 2 不同特征参数的预测成功率

Fig.2 Prediction accuracy of different feature parameters

2.2.2 融合特征预测结果

在选取单个特征对亚高尔基体蛋白质定位预测的基础上,进一步对单个特征进行融合,利用融合特征对亚高尔基体蛋白质定位预测,预测结果表明:GO 特征与其它特征的融合结果好于没有 GO 特征的融合。表 4 是 GO 特征与其它一些特征信息融合后的预测结果。

表 4 融合特征参数的预测结果

Table 4 Prediction results of fusion feature parameters

特征	Sn/%	Sp/%	Acc/%	MCC
GO+AAC	95.05	57.33	87.43	0.583
GO+DC	97.88	41.33	86.31	0.524
GO+APAAC	96.11	49.33	86.59	0.543
GO+CTF	96.47	50.67	87.16	0.563
GO+acACS	92.93	64.0	87.15	0.591
GO+DC+APAAC	95.41	53.33	86.87	0.558
GO+AAC+APAAC	97.17	49.33	87.43	0.571
GO+AAC+DC+APAAC	95.41	52.0	86.59	0.548
GO+AAC+acACS+APAAC	96.82	48.0	86.87	0.550
GO+AAC+acACS+APAAC+CTF	95.05	54.67	86.87	0.561

图 3表示融合特征参数的预测准确率 Acc,由图 3 可知,单肽组分信息和基因本体注释信息融合;单肽组分信息、两亲性伪氨基酸组分信息和基因本体注释信息融合,这两组融合特征参数的 Acc 最高。由表 4 可以看出,融合特征对区分 cis-Golgi 蛋白质和 trans-Golgi 蛋白质有一定的作用。两种融合特征信息中,单肽组分信息和基因本体注释信息融合后预测成功率提高到 87.43%,其它特征信息与基因本体注释信息融合以后预测成功率没有提高,还有所下降,出现这种结果的原因可能是:首先,融合特征信息以后,特征信息维数过大,数据冗余,影响预测结果;其次,特征信息融合以后一些蕴含亚高尔基体位置的关键信息可能被覆盖或者丢失,不能有效识别亚高尔基体蛋白质。三种特征信息融合中,单肽组分信息、两亲性伪氨基酸组分信息和基因本体注释信息融合的预测成功率提高到 87.43%,以上结果说明两种和三种特征信息适当的融合对亚高尔基体蛋白质的识别有一定的作用。由表 3 和表 4 可以看出,选取单个特征和融合特征对亚高尔基体蛋白质的预测,敏感性 Sn 都较高,这表明本文的预测方法对 trans-Golgi 蛋白质的识别具有一定的优势,一些特征信息融合以后,特异性 Sp 和 Matthew 相关系数 MCC 比单个特征信息会有一些提高,当基因本体(GO)注释信息分别与氨基酸单肽组分信息(AAC)和平均化学位移(acACS)融合以后 Sp 和 MCC 有一定的提升,说明这两类融合特征对 cis-Golgi 蛋白质识别的准确率有一些提升作用,其它融合特征的特异性 Sp 和 Matthew 相关系数 MCC 并没有提高。

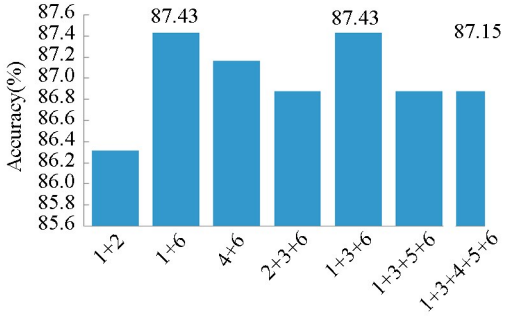

图 3　融合特征参数的预测成功率

Fig.3　Prediction accuracy of fusion feature parameters

注:其中 1 代表 AAC,2 代表 DC,3 代表 APAAC,4 代表 CTF,5 代表 acACS,6 代表 GO.

本文对亚高尔基体蛋白质的预测结果相比于前人的预测结果,总体预测成功率略有提高,但是由于本文的两类亚高尔基体数据集是新建立的数据集,新建的数据集蛋白质数量有所增加,本文是针对新建立的数据集进行预测的,所以预测结果与前人的预测结果不具可比性。

3　结　论

正确识别亚高尔基体蛋白质的类型对治疗由高尔基体蛋白质功能紊乱而引发的疾病有一定的作用。本文建立了两类亚高尔基体蛋白质数据集,选取多种特征信息,利用支持向量机算法,基于 5-折交叉检验对该数据集进行预测,预测结果表明:1)选取的 6 种特征信息中基因本体 GO 注释信息对亚高尔基体蛋白质定位有较高识别率,本文可以有效识别 trans-Golgi 蛋白质,但是对于 cis-Golgi 蛋白质的识别效果较差;2)基因本体 GO 注释信息与其它特征信息进行恰当融合后对预测结果有一定的提升作用。

参考文献(References)

[1]李明俊.利用多信息融合方法预测蛋白质亚核定位[D].呼和浩特:内蒙古农业大学,2019. DOI:10.27229/d.cnki.gnmnu.2019.000115.
LI Mingjun. Prediction of protein subnuclear localization based on different features[D]. Hohhot:Inner Mongolia Agricultural University, 2019. DOI: 10.27229/d. cnki. gnmnu. 2019.000115.

[2]张松,黄波,夏学峰,等.蛋白质亚细胞定位的生物信息学研究[J].生物化学与生物物理进展,2007,34(6):573-579. DOI: 10.3321/j.issn:1000-3282.2007.06.004.
ZHANG Song, HUANG Bo, XIA Xuefeng, et al. Bioinformatics research in subcellular localization of protein[J]. Progress in Biochemistry and Biophysics, 2007, 34(6): 573-579. DOI: 10.3321/j.issn:1000-3282.2007.06.004.

[3]FUJITA Y, OHAMA E, TAKATAMA M, et al. Fragmentation of Golgi apparatus of nigral neurons with alpha-synuclein-positive inclusions in patients with Parkinson's disease[J]. Acta Neuropathologica, 2006, 112(3): 261-265. DOI: 10.1007/s00401-006-0114-4.

[4]GONATAS N K, GONATAS J O, STIEBER A. The involvement of the Golgi apparatus in the pathogenesis of amyotrophic lateral sclerosis, Alzheimer's disease, and ricin intoxication[J]. Histochemistry & Cell Biology, 1998, 109(5/6): 591-600. DOI:10.1007/s004180050257.

[5]苗芳芳,范隆,王天龙.高尔基体与神经退行性疾病[J].北京医学, 2017, 39(3): 294-296. DOI: CNKI:SUN:BJYX.0.2017-03-024.
MIAO Fangfang, FAN Long, WANG Tianlong. Golgi apparatus and neurodegenerative diseases[J]. Beijing Medicine, 2017, 39(3): 294-296. DOI: CNKI:SUN:BJYX.0.2017-03-024.

[6]张瑜,周文胜,王佳. 高尔基体与神经退行性疾病研究进展[J].中风与神经疾病杂志,2017,34(5):474-477. DOI: CNKI:SUN:ZFSJ.0.2017-05-026.
ZHANG Yu, ZHOU Wensheng, WANG Jia. Research progress of Golgi body and neurodegenerative diseases[J]. Journal of Stroke and Neurological Diseases,2017,34(5): 474-477.DOI:CNKI:SUN:ZFS J.0.2017-05-026.

[7]DIJK A, BRAAK C. Predicting sub-Golgi localization of type II membrane proteins [J]. Bioinformatics, 2008, 24(16):1779-1786. DOI: 10.1093/bioinformatics/btn309.

[8]DING H, LIU L, GAO F B, et al. Identify Golgi protein types with modified Mahalanobis discriminant algorithm and pseudo amino acid composition [J]. Protein and Peptide Letters, 2011, 18 (1): 58 - 63. DOI: 10. 2174/092986611794328708.

[9]DING H, GUO S H, DENG E Z, et al. Prediction of Golgi-resident protein types by using feature selection technique[J]. Chemometrics and Intelligent Laboratory Systems,2013,124(6):9-13. DOI: 10.1016/j.chemolab.2013.03.005.

[10]JIAO Y S, DU P F. Predicting Golgi-resident protein types using pseudo amino acid compositions: Approaches with positional specific physicochemical properties[J]. Journal of Theoretical Biology,2016,391(5)35-42. DOI: 10.1016/j.jtbi.2015.11.009.

[11]YANG R T, ZHANG C J, GAO R, et al. A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data [J]. IJMS, 2016,17(2):218. DOI:10.3390/ijms17020218.

[12]AHMAD J, JAVED F, HAYAT M. Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods[J]. Artificial Intelligence in Medicine, 2017,78(5):14-22. DOI: 10.1016/j.artmed.2017.05.001.

[13]RAHMAN M S, RAHMAN M K, KAYKOBAD M, et al. isGPT: An optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection[J]. Artificial Intelligence in Medicine, 2018, 84(5):90-100. DOI:10.1016/j.artmed.2017.11.003.

[14]张蕾.基于序列和结构信息预测蛋白质亚高尔基体定位[D].呼和浩特:内蒙古大学,2018. DOI: 10.7666/d.D01535 807.
ZHANG Lei. Prediction of protein sub-Golgi localization based on sequence and structure information[D]. Hohhot: Inner Mongolia University, 2018. DOI: 10. 7666/d.D01535807.

[15]LV Z B, JIN S S, DING H, et al. A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features[J]. Frontiers in Bioengineering and Biotechnology,2019,7:215. DOI: 10.3389/fbioe.2019.00215.

[16]ZHAO W, LI G P, WANG J, et al. Predicting protein sub-Golgi locations by combining functional domain enrichment scores with pseudo-amino acid compositions [J]. Journal of Theoretical Biology, 2019,473:38-43. DOI:10.1016/j.jtbi.2019.04.025.

[17]CUI Q Y, CAO Y, BAO W Z, et al. SubRF_Seq: Identification of Sub-Golgi protein types with random forest with partial sequence information[J]. Scientific Programming, 2020,2020(9):1-7. DOI: 10.1155/2020/8862468.

[18]罗林波,陈绮.氨基酸序列特征提取方法研究[J].计算机技术与发展,2010,20(2):206-208. DOI:10.3969 /j.issn.1673-629X.2010.02.054.
LUO Linbo, CHEN Qi. Research on feature extraction method of amino acid sequence[J]. Computer Technology and Development,2010,20(2):206-208. DOI:10.3969/j.issn.1673-629X. 2010.02.054.

[19]AHMAD K, WARIS M, HAYAT M. Prediction of protein submitochondrial locations by incorporating dipeptide composition into chou's general pseudo amino acid composition [J]. The Journal of Membrane Biology, 2016, 249(3):293-304. DOI:10.1007/s00232-015-9868-8.

[20]CHOU K C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes [J]. Bioinformatics, 2005, 21 (1): 10 - 19. DOI: 10. 1093/bioinformatics/bth466.

[21]SHEN J W, ZHANG J, LUO X M, et al. Predicting protein-protein interactions based only on sequences information[J]. Proceedings of the National Academy of Sciences of The United States of America,2007,104(11): 4337-4341. DOI:10.1073/pnas.0607879104.

[22]SIBLEY A B, COSMAN M, KRISHNAN V V. An empirical correlation between secondary structure content and averaged chemical shifts in proteins[J].Biophysical Journal, 2003,84:1223-1227.DOI: 10.1016/ S0006-3495(03) 74937-6.

[23]高晓伟.革兰氏阳性菌蛋白质亚细胞定位的特征提取及预测算法研究[D].呼和浩特:内蒙古农业大学,2020. DOI:10.27229/d.cnki.gnmnu.2020.000810.
GAO Xiaowei. Study on feature extraction and prediction algorithm for subcellular localization of gram-positive bacterial protein[D]. Hohhot: Inner Mongolia Agricultural University,2020. DOI:10.27229/d.cnki.gnmnu.2020.000810.

[24]GIROSI F. An equivalence between sparse approximation and support vector machines[J].Neural Computation,1998, 10(6):1445-1480. DOI:10.1162/089976698300017269.

[25]JING X Y, LI F M. Identifying heat shock protein families from imbalanced data by using combined features [J]. Computational and Mathematical Methods in Medicine, 2020,2020:8894478. DOI:10.1155/ 2020/8894478.