(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104331642 B
(45)授权公告日 2017.04.12

(21)申请号 201410588610.8

(22)申请日 2014.10.28

(65)同一申请的已公布的文献号
申请公布号 CN 104331642 A

(43)申请公布日 2015.02.04

(73)专利权人 山东大学
地址 250061 山东省济南市历下区经十路17923号

(72)发明人 张承进 杨润涛 高瑞 张丽娜

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 张勇

(51)Int.Cl.
G06F 19/24(2011.01)
G06F 19/18(2011.01)

(56)对比文件
CN 101145171 A,2008.03.19,
CN 102012977 A,2011.04.13,
WO 2013190084 A1,2013.12.27,
宋佳.机器学习方法在生物序列分析中的应用.《万方数据库论文在线出版》.2014,
钮冰.基于集成学习算法的若干生物信息学问题研究.《中国博士学位论文全文数据库-基础科学辑》.2010,(第05期),
晏春等.基于支持向量机的生物序列分析.《计算机仿真》.2006,第23卷(第9期),

审查员 贺馨

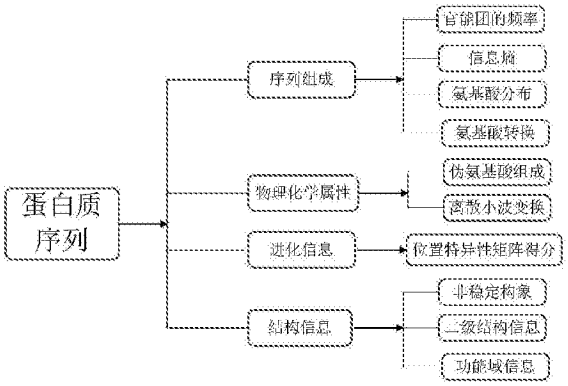
权利要求书2页 说明书10页 附图4页

(54)发明名称

用于识别细胞外基质蛋白的集成学习方法

(57)摘要

本发明公开了用于识别细胞外基质蛋白的集成学习方法,数据集建立:建立细胞外基质ECM蛋白序列的训练样本集和独立测试样本集;将训练样本集中的蛋白质序列映射成数值特征向量;采用信息增益率—增量特征选择方法挑选出相对有效的特征子集,采用集成学习的方法建立集成分类器模型,以解决数据集不平衡的问题;将独立测试样本集映射成数值特征向量,基于集成分类器模型的预测结果,采用多数表决方法得到测试样本的类别,最终利用所有测试样本的预测结果评价预测系统的性能;本发明开发了用于细胞外基质蛋白识别的网络服务器系统。用户无需理解细胞外基质蛋白识别的具体执行过程,只需输入待预测的蛋白质序列,即可得到预测结果。



CN 104331642 B

1.用于识别细胞外基质蛋白的集成学习方法,其特征是,包括以下步骤:

步骤一:数据集建立:建立细胞外基质ECM蛋白序列的训练样本集和独立测试样本集;

步骤二:基于序列组成、物理化学属性、进化信息及结构信息,将训练样本集中的蛋白质序列映射成数值特征向量;

步骤三:为降低计算复杂度和减少特征的冗余性,采用信息增益率—增量特征选择方法挑选出相对有效的特征子集,以提高评估训练样本集的预测性能;

步骤四:采用集成学习的方法建立集成分类器模型,以解决数据集不平衡的问题;

步骤五:将独立测试样本集按步骤二的方法映射成数值特征向量,基于集成分类器模型的预测结果,采用多数表决方法得到测试样本的类别,最终利用所有测试样本的预测结果评价预测系统的性能;

步骤六:利用用于细胞外基质蛋白识别的网络服务器系统,进行在线识别细胞外基质蛋白;

所述步骤四中的分类器模型为随机森林,随机森林通过重采样技术,随机生成训练样本用于训练多个决策树,基于多数表决的方法,独立测试样本的最终预测结果由决策树输出类别的众数而定;

所述步骤二中,所述蛋白质序列映射成数值特征向量的方法为:基于序列组成官能团的频率的特征建立策略;基于序列组成信息熵的特征建立策略;基于序列组成氨基酸分布的特征建立策略;基于序列组成氨基酸转换的特征建立策略;基于物理化学属性伪氨基酸组成的特征建立策略;基于物理化学属性离散小波变换的特征建立策略;基于进化信息的特征建立策略;基于进化信息非稳定构象的特征建立策略;基于进化信息二级结构信息的特征建立策略;基于进化信息功能域信息的特征建立策略;

所述步骤三中,采用信息增益率—增量特征选择方法挑选出相对有效的特征子集,具体为:利用增量特征选择方法获取最优特征子集,增量特征选择方法从空特征集合开始,按特征的排序从高到低逐一加入到特征集合;每一次加入一个特征,都会产生一个新的特征子集,具有高均衡准确率和低维数的特征子集将被作为预测系统的最终输入特征向量。

2.如权利要求1所述的用于识别细胞外基质蛋白的集成学习方法,其特征是,所述步骤一中训练样本集含有410个后生动物ECM蛋白和4464个后生动物非ECM蛋白;独立测试样本集则包括85个人类ECM蛋白和130个人类非ECM蛋白。

3.如权利要求1所述的用于识别细胞外基质蛋白的集成学习方法,其特征是,所述步骤五中评价预测系统的性能指标分别为敏感性sensitivity、特异性specificity、准确率accuracy、均衡准确率balanced accuracy;上述评价指标定义分别如下:

$$S_n = \frac{TP}{TP + FN},$$

$$S_p = \frac{TN}{TP + FP},$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN},$$

$$BAcc = \frac{1}{2}(S_n + S_p).$$

其中，TP、FN、TN和FP分别为真阳性True Positive、假阴性False Negative、真阴性True Negative和假阳性False Positive。

4.如权利要求1所述的用于识别细胞外基质蛋白的集成学习方法，其特征是，评估训练样本集的预测性能的方法为10-交叉验证方法，训练样本集的正负样本集分别随机地分为样本数量相同的10组数据子集，在这生成的20组数据子集中，正负样本集的各一组数据子集用于训练，剩余的数据子集用于测试，每次用于训练的数据子集保证不同，上述过程循环重复10次。

5.如权利要求4所述的用于识别细胞外基质蛋白的集成学习方法，其特征是，对于每一次循环过程，其执行流程包括如下步骤：

S1：训练样本集中负样本的个数大约为正样本的11倍，负样本训练集通过欠采样方法分为样本数量几乎相同的11组数据子集，每一组数据子集与正样本训练集构成训练子集，通过上述欠采样过程，可得到11个训练子集；

S2：分别用S1得到的11个训练子集训练随机森林分类器，所获取的11个随机森林分类器组成集成分类器，测试样本集用于评估集成分类器的性能，基于集成分类器，测试样本的最终预测类别通过多数表决的方法获得。

用于识别细胞外基质蛋白的集成学习方法

技术领域

[0001] 本发明涉及蛋白质功能属性识别领域，具体为一种用于识别细胞外基质蛋白的集成学习方法。

背景技术

[0002] 细胞外基质(Extracellular Matrix,ECM)是细胞和组织赖以生存的微环境，在细胞行为和组织特性的调控中发挥重要作用。ECM强大的生物学功能归因于ECM蛋白的多样性。ECM蛋白的组成和动态变化对细胞的增殖、分化、迁移，组织的形态发生、分化等生命现象具有全方位的影响。同时，ECM蛋白的功能紊乱可导致众多疾病。蛋白聚糖和胶原是ECM蛋白的主要组成成分。其中，蛋白聚糖调控组织修复、肿瘤生长、细胞粘附、增殖和迁移等生理活动；胶原蛋白广泛应用于骨组织工程，并调节细胞粘附、迁移，指导组织发育。ECM蛋白质的准确识别将有助于理解上述生物过程的潜在机制，并为基于ECM蛋白的生物材料设计和药物开发提供重要的线索。

[0003] 近二十多年来，生命科学快速发展的最重要特征是生物学数据量的剧增。如何处理、分析和解释这些生物学数据成为众多学者关注的问题。其中，生物大分子序列的功能属性识别问题已成为生物信息学领域的重要研究课题，由于实验测定方法昂贵而且周期长，模式识别方法已成为主流方法。近年来，研究人员尝试应用机器学习方法识别细胞外基质蛋白。2010年，Juan J等建立了ECM蛋白的预测系统ECMPP，此方法引入了5种新特征，包括分子量、序列长度、重复残基、重复结构域、重复三联体glycine-x-y(Jung J,Ryu T,Hwang Y,Lee E,Lee D.(2010)Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics.Journal of computational Biology 17:97-105)。2013年，Kandaswamy KK等开发了预测ECM蛋白的网络服务器ECMPRED，该方法所提取的特征来自于蛋白质序列中官能团的频率和氨基酸的物理化学性质(Kandaswamy KK,Pugalenthi G,Kalies KU,Hartmann E,Martinetz T.(2013)EcmPred:prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection.Journal of Theoretical Biology 317:377-383)。然而，对蛋白质功能属性预测非常重要的序列顺序信息和结构信息，上述两种方法均未考虑。而且，现有方法也没有解决数据集不平衡的问题(ECM蛋白的样本个数远远小于非ECM蛋白的样本个数)，导致绝大多数样本被预测为非ECM蛋白，极大地限制了分类器的性能。

发明内容

[0004] 为解决现有技术存在的不足，本发明公开了用于识别细胞外基质蛋白的集成学习方法，目的在于解决数据集的不平衡问题，同时综合多种序列特征信息，以平衡和提高细胞外基质蛋白正负样本的预测精度。

[0005] 为实现上述目的，本发明的具体方案如下：

[0006] 用于识别细胞外基质蛋白的集成学习方法，包括以下步骤：

[0007] 步骤一：数据集建立：建立细胞外基质ECM蛋白序列的训练样本集和独立测试样本集；

[0008] 步骤二：基于序列组成、物理化学属性、进化信息及结构信息，将训练样本集中的蛋白质序列映射成数值特征向量；

[0009] 步骤三：为降低计算复杂度和减少特征的冗余性，采用信息增益率—增量特征选择方法挑选出相对有效的特征子集，以提高评估训练样本集的预测性能；

[0010] 步骤四：采用集成学习的方法建立集成分类器模型，以解决数据集不平衡的问题；

[0011] 步骤五：将独立测试样本集按步骤二的方法映射成数值特征向量，基于集成分类器模型的预测结果，采用多数表决方法得到测试样本的类别，最终利用所有独立测试样本的预测结果评价预测系统的性能；

[0012] 步骤六：利用用于细胞外基质蛋白识别的网络服务器系统，进行在线识别细胞外基质蛋白。

[0013] 所述步骤一中训练样本集含有410个后生动物ECM蛋白和4464个后生动物非ECM蛋白；独立测试样本集则包括85个人类ECM蛋白和130个人类非ECM蛋白。

[0014] 所述步骤二中，所述蛋白质序列映射成数值特征向量的方法为：基于序列组成官能团的频率的特征建立策略；基于序列组成信息熵的特征建立策略；基于序列组成氨基酸分布的特征建立策略；基于序列组成氨基酸转换的特征建立策略；基于物理化学属性伪氨基酸组成的特征建立策略；基于物理化学属性离散小波变换的特征建立策略；基于进化信息的特征建立策略；基于进化信息非稳定构象的特征建立策略；基于进化信息二级结构信息的特征建立策略；基于进化信息功能域信息的特征建立策略。

[0015] 所述步骤三中，采用信息增益率—增量特征选择方法挑选出相对有效的特征子集，具体为：利用增量特征选择方法获取最优特征子集，增量特征选择方法从空特征集合开始，按特征的排序从高到低逐一加入到特征集合；每一次加入一个特征，都会产生一个新的特征子集，具有高均衡准确率和低维数的特征子集将被作为预测系统的最终输入特征向量。

[0016] 所述步骤四中的分类器模型为随机森林，随机森林通过重采样技术，随机生成训练样本用于训练多个决策树，基于多数表决的方法，独立测试样本的最终预测结果由决策树输出类别的众数而定。

[0017] 所述步骤五中评价预测系统的性能指标分别为敏感性sensitivity、特异性specificity、准确率accuracy、均衡准确率balanced accuracy；上述评价指标定义分别如下：

[0018] $$S_n = \frac{TP}{TP + FN},$$

[0019] $$S_p = \frac{TN}{TP + FP},$$

[0020] $$Acc = \frac{TP + TN}{TP + FP + TN + FN},$$

[0021] $$BAcc = \frac{1}{2}(S_n + S_p).$$

[0022] 其中，TP、FN、TN和FP分别为真阳性True Positive、假阴性False Negative、真阴性True Negative和假阳性False Positive。

[0023] 评估训练样本集的预测性能的方法为10-交叉验证方法，训练样本集的正负样本集分别随机地分为样本数量相同的10组数据子集，在这生成的20组数据子集中，正负样本集的各一组数据子集用于训练，剩余的数据子集用于测试，每次用于训练的数据子集保证不同，上述过程循环重复10次。

[0024] 对于每一次循环过程，其执行流程包括如下步骤：

[0025] S1：训练样本集中负样本的个数大约为正样本的11倍，负样本训练集通过欠采样方法分为样本数量几乎相同的11组数据子集。每一组数据子集与正样本训练集构成训练子集，通过上述欠采样过程，可得到11个训练子集；

[0026] S2：分别用S1得到的11个训练子集训练随机森林分类器，所获取的11个随机森林分类器组成集成分类器，测试样本集用于评估集成分类器的性能，基于集成分类器，测试样本的最终预测类别通过多数表决的方法获得。

[0027] 数据集建立的具体过程：Kandaswamy KK等给出的445个ECM蛋白和4486个非ECM蛋白用于训练样本集的构造。与此同时，利用人类蛋白质组建立独立测试集。人类ECM蛋白(正样本)从文献(Cromar GL,Xiong X,Chautard E,Ricard-Blum S,Parkinson J.(2012) Toward a systems level view of the ECM and related proteins:a framework for the systematic definition and analysis of biological systems.Proteins 80: 1522-1544)附件3中提取，人类非ECM蛋白(负样本)则通过文献(Li L,Zhang Y,Zou L,Li C,Yu B,et al.(2012)An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity.PLoS ONE 7:e31057)中的Hum3681数据集进行搜集。Hum3681数据集包含14个亚细胞位置的蛋白质序列，排除细胞外基质，分别从剩余亚细胞位置的蛋白质序列中随机选择10条序列作为独立测试集的负样本。为获得优质的数据，长度小于50或大于3000的蛋白质序列将被筛选掉；另外，舍弃从Unipro数据库删除、含有非天然氨基酸的蛋白质序列。最终，训练样本集含有410个后生动物ECM蛋白和4464个后生动物非ECM蛋白；独立测试样本集则包括85个人类ECM蛋白和130个人类非ECM蛋白。

[0028] 灵敏性和特异性分别反映了正样本和负样本的预测准确率；准确率则是所有样本的预测准确率。对于不平衡数据集的分类，通常会出现灵敏性非常低、准确率仍很高的情形。因此，准确率指标不能有效地评估不平衡数据集分类问题。通常希望预测系统同时具有高灵敏性和高特异性的特点。为此，本发明引入性能评估的主要指标——均衡准确率，此指标定义为灵敏性和特异性的平均值。

[0029] 随机森林具有高精度、训练速度快、能够处理高维数据等优点。本发明利用数据挖掘工具WEKA中带有默认参数的随机森林算法实施分类。

[0030] 本发明的有益效果：

[0031] 1.在蛋白质特征建立阶段，本发明综合考虑了蛋白质序列的各方面信息，包括序列组成、物理化学属性、进化信息和结构信息。这种全方位的特征建立策略将使各类特征之

间形成互补关系，有利于分类器性能的提高。

[0032] 2.通常原始的特征集合含有很多冗余信息和噪声，本发明采用信息增益率—增量特征选择方法排除冗余特征，以减少“维数灾难”和提高分类器性能。

[0033] 3.训练数据集中正样本的个数远远小于负样本的个数，这种数据集不平衡的问题将导致预测系统的敏感性很低。本发现利用集成学习的方法解决了此问题，得到了具有高敏感性和高特异性的细胞外基质蛋白预测系统。

[0034] 4.本发明开发了用于细胞外基质蛋白识别的网络服务器系统。用户无需理解细胞外基质蛋白识别的具体执行过程，只需输入待预测的蛋白质序列，即可得到预测结果。

附图说明

[0035] 图1蛋白质序列特征建立策略图；

[0036] 图2信息增益率—增量特征选择方法过程；

[0037] 图3增量特征选择方法曲线图；

[0038] 图4不平衡数据集预测性能变化曲线；

[0039] 图5集成学习方法执行流程；

[0040] 图6 ICEMP网络服务器的主页；

[0041] 图7 ICEMP网络服务器的预测结果页面。

具体实施方式：

[0042] 下面结合附图对本发明进行详细说明：

[0043] 为建立用于蛋白质功能属性识别的计算方法，首先应将蛋白质序列表示为数值特征向量。图1给出了本发明的特征建立策略。基于序列组成、物理化学属性、进化信息和结构信息，本发明采取10种特征建立方法将蛋白质序列映射成维数为315的数值特征向量。下面逐一阐明每一种特征建立策略。

[0044] 1.基于序列组成的特征建立策略

[0045] (I)官能团的频率

[0046] 氨基酸的侧链在蛋白质的结构折叠和稳定过程中扮演重要角色。基于侧链的化学基团，本发明将20种天然氨基酸按官能团类别分成10组，分别为苯基(F/W/Y)，羧基(D/E)，咪唑(H)，伯胺(K)，胍基(R)，硫醇(C)，硫(M)，氨基(Q/N)，羟基(S/T)和非极性(A/G/I/L/V/P)。分别计算这10组官能团在蛋白质序列中出现的频率。

[0047] (II)信息熵

[0048] 在自然选择下，蛋白质的氨基酸组成可看作一不确定性系统。在信息理论中，熵可以合理地描述随机变量的不确定性。作为信息理论中最重要的一个指标，香农熵可以表示为

[0049] $$H(x)=-\sum_{i=1}^{n}P_i\log_2 P_i.$$

[0050] 根据如上公式，分别计算氨基酸组成和二肽组成的香农熵。其中，P_i(i＝1,2,…,n)分别为20种天然氨基酸和400种二肽在蛋白质序列中出现的频率。

[0051] (III)氨基酸分布

[0052] 蛋白质序列中每一种天然氨基酸的个数记为N_i(i=1,2,…,20)。D_j^i为蛋白质序列中的第j个氨基酸i与第一个氨基酸i的距离。则氨基酸i的分布定义为

[0053] $$D_i=\sum_{j=1}^{N_i}\frac{\left(D_j^i-AD_j^i\right)^2}{N_i},$$

[0054] 其中$AD_j^i=\frac{1}{N_i}\sum_{j=1}^{N_i}D_j^i$。

[0055] (IV)氨基酸转换

[0056] 为避免丢失蛋白质序列的顺序信息,采取氨基酸的转换特征刻画蛋白质序列,其求取公式为

[0057] $$T_{\alpha_i,\alpha_j}=\frac{N_{\alpha_i,\alpha_j}+N_{\alpha_j,\alpha_i}}{L},$$

[0058] 其中i,j∈{1,2,…,10},且i≠j。α_i表示10种官能团中的一种,$N_{\alpha_i\alpha_j}$为二肽“$\alpha_i\alpha_j$”在蛋白质序列中出现的次数,L为蛋白质序列的长度。

[0059] 2.基于物理化学属性的特征建立策略

[0060] (I)伪氨基酸组成

[0061] 蛋白质结构、功能的特异性及多样性在很大程度上与氨基酸的物理化学属性相关。伪氨基酸组成结合了氨基酸的物理化学属性和蛋白质序列的顺序信息,已广泛应用于蛋白质功能属性的识别问题中。有关伪氨基酸组成的模型众多,本发明将采用文献(Afridi TH,Khan A,Lee YS.(2012)Mito-GSAAC:mitochondria prediction using genetic ensemble classifier and split amino acid composition.Amino Acids 42:1443-1454)中的模型用于提取伪氨基酸组成特征。令参数η=20,则从此模型中可以得到40个特征。

[0062] 基于以下原因,本发明将考虑4种物理化学属性计算伪氨基酸组成模型,分别为疏水性、柔韧性、净电荷、和平均接触表面积。(i)疏水作用被认为是影响蛋白质结构的最重要的因素;(ii)作为一类ECM蛋白,胶原蛋白分子的柔韧性对于细胞行为的调控至关重要;(iii)带电氨基酸更倾向于形成氢键,有利于ECM蛋白质与溶剂分子发生相互作用;(iv)氨基酸的平均接触表面积与蛋白质翻译后修饰行为密切相关,可能是ECM形成动态网络的驱动力。

[0063] (II)离散小波变换

[0064] 离散小波变换可以同时在时域和频域上对信号进行分析,因此在基因组序列分析、蛋白质结构预测、基因表达数据分析等研究中得到了广泛的应用。通过离散小波变换,原始信号可以分解为信号的近似值和信号的细节值。在小波分析中,近似值是大的缩放因子产生的系数,表示信号的低频分量;而细节值是小的缩放因子产生的系数,表示信号的高频分量。根据数据分析的需要,可以对原始信号进行多级分解,得到每一个子带信号的近似值和细节值。本发明将首先利用疏水性、柔韧性和平均接触表面积分别将蛋白质序列转换成数值序列,然后对得到的数值序列实施小波变换,以提取蛋白质序列物理化学属性的频谱特征。

[0065] 本发明通过离散小波变换所建立的特征如下:(i)原始信号的平均值和方差;(ii)

每一个子带小波系数的最大值、最小值、平均值及方差。在这里，选择“Db4”作为小波函数，信号的分解级数设为4。则对于每个蛋白质序列，可得到42个物理化学属性的频谱特征。

[0066] 3.基于进化信息的特征建立策略

[0067] 蛋白质的生物学功能通常体现在其序列的进化保守性上，越来越多的证据表明进化信息对于蛋白质的结构和功能预测至关重要。本发明将利用位置特异性得分矩阵(Position Specific Scoring Matrix，PSSM)提取蛋白质序列的进化信息。通过“PSI-BLAST”网络服务器3次迭代，序列长度为L的蛋白质可生成维数为L×20的PSSM矩阵。

[0068] $$P_{PSSM}=\begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to j} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to j} & \cdots & E_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to j} & \cdots & E_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to j} & \cdots & E_{L\to20} \end{bmatrix},$$

[0069] 其中，$E_{i\to j}$表示在进化的过程中蛋白质序列第i位置处的氨基酸突变成氨基酸类型j的得分。PSSM通过如下的s型函数标准化。

[0070] $$f(x)=\frac{1}{1+e^{-x}},$$

[0071] 其中x为原始的PSSM值。基于进化信息，蛋白质序列可以表示为

[0072] $$F_{PSSM}=\left[\theta_1^1,\theta_2^1,\cdots,\theta_{20}^1,\theta_1^2,\theta_2^2,\cdots,\theta_{20}^2,\cdots,\theta_1^\lambda,\theta_2^\lambda,\cdots,\theta_{20}^\lambda\right]$$

[0073] $$\theta_i^\lambda=\frac{1}{L-\lambda}\sum_{j=1}^{L-\lambda}\left(E_{j\to i}-E_{j+\lambda\to i}\right)^2,\ i=1,2,\cdots,20,\ 0<\lambda<L.$$

[0074] 选取λ值为4，则从蛋白质序列的进化信息中提取的特征总数为80。

[0075] 4.基于结构信息的特征建立策略

[0076] (I)非稳定构象

[0077] 非稳定构象是指不能折叠形成稳定三维结构的蛋白质区域。非稳定构象区域在转录调控、细胞信号传导、翻译后修饰等各种信号与调控路径中发挥着重要作用。通过非稳定构象预测器“VSL2”，可以计算给定蛋白质序列的每一个氨基酸的非稳定构象得分。下面的8个数值特征将用来编码蛋白质序列。(i)所有氨基酸非稳定构象得分的平均值和方差(2个特征)。(ii)稳定构象区域、非稳定构象区域的个数(2个特征)。(iii)稳定构象区域、非稳定构象区域的最小、最大长度(4个特征)。

[0078] (II)二级结构信息

[0079] 蛋白质的二级结构指多肽链通过氢键沿一定方向盘绕、折叠而形成的构象。多个二级结构单元在空间排列形成三维结构，其在很大程度上决定了蛋白质的功能。二级结构构象主要包括α-螺旋、β-折叠和无规则卷曲。本发明采用二级结构预测工具“PSIPRED”将蛋白质序列映射成二级结构序列，然后从中提取如下51个数值特征。(i)3种二级结构构象在蛋白质序列中出现的频率(3个特征)；(ii)3种二级结构构象的分布(3个特征)；(iii)3种二级结构构象区域的个数(3个特征)；(iv)3种二级结构构象区域长度的最小值、最大值、平均值和方差(12个特征)；(v)10种官能团在三种二级结构构象的频率(30个特征)。

[0080] (III)功能域信息

[0081] 功能域是蛋白质分子中具有特异结构和独立功能的区域，执行多种生物学功能。在同一细胞器的蛋白质通常拥有相同的功能域。因此，本发明将从蛋白质功能域信息中提取特征。首先，从Intepro数据库中获取训练数据集中每一个ECM蛋白质的功能域组成。然后，从获得的所有功能域中挑选出不少于25个ECM蛋白共有的功能域，这些功能域用于后续的特征提取。经过以上两步，最终获得了17种功能域。这17种功能域被表示成维数为17的二进制向量，如果某种功能域存在于蛋白质序列中，则令其对应的二进制特征值为1，否则为0。这样就从蛋白质的功能域信息中提取了17个数值特征。

[0082] 经过以上特征建立方法，蛋白质序列被转换成了维数为315的数值特征向量。然而，原始的特征集合中通常含有冗余信息和噪声，这将导致预测性能降低和维数灾难。因此，对原始特征集合进行特征选择至关重要，本发明将采取信息增益率—增量特征选择方法挑选出相对有效的特征子集，以提高预测性能。结合图2给出的特征选择过程，下面将详细说明信息增益率—增量特征选择方法的原理。

[0083] 信息增益率能够准确地刻画特征与预测类别的相关性。在本发明中，类别C的信息熵定义为

[0084] $$H(C)=-\sum_{j=1}^{2}P(C_j)\log_2 P(C_j),$$

[0085] 其中$P(C_j)$为类别C_j(ECM蛋白或非ECM蛋白)在训练数据集中的比例。

[0086] 特征F_i($i\in\{1,2,\cdots,315\}$)的特征值集合记为$S_i=\{V_i^1,V_i^2,\cdots,V_i^{n_i}\}$。特征$F_i$的信息熵表示为

[0087] $$H(F_i)=-\sum_{j=1}^{n_i}P(V_i^j)\log_2 P(V_i^j).$$

[0088] 给定特征F_i，类别C的条件信息熵定义为

[0089] $$H(C|F_i)=-\sum_{j=1}^{n_i}P(V_i^j)\sum_{k=1}^{2}P(C_k|V_i^j)\log_2 P(C_k|V_i^j).$$

[0090] 则特征F_i的信息增益率为

[0091] $$IGR(F_i)=\frac{H(C)-H(C|F_i)}{H(F_i)}.$$

[0092] 根据信息增益率测度，若$IGR(F_i)>IGR(F_j)$，则与特征F_j相比，特征F_i与类别C更相关，即特征F_i对分类更重要。基于信息增益率，可以对特征的重要性进行排序。

[0093] 本发明利用增量特征选择方法获取最优特征子集。增量特征选择方法从空特征集合开始，按特征的排序从高到低逐一加入到特征集合；每一次加入一个特征，都会产生一个新的特征子集。具有高均衡准确率和低维数的特征子集将被作为预测系统的最终输入特征向量。

[0094] 增量特征选择方法的结果如图3所示，图3呈现了均衡准确率和特征子集的关系。从图3中可以看出，当特征子集维数为289时，均衡准确率达到了最大值0.8645。而特征子集维数为102时，均衡准确率达到了0.8635，仅仅比最大值小0.001。为避免维数的灾难，此102个特征作为最终的最优特征子集用于细胞外基质蛋白的识别。

[0095] 从训练数据集中可以看出，ECM蛋白的个数远远少于非ECM蛋白的个数。为分析这种不平衡数据集对于预测性能的影响，本发明通过随机从训练数据集中选取负样本，再加上训练数据集的全部正样本，构成了10组训练数据子集。这10组训练数据子集中正负样本个数之比分别为1:1到1:10。利用10-交叉验证，图4给出了这10组训练数据子集预测性能的变化曲线。

[0096] 如图4所示，随着负样本的增加，特异性逐渐提高。与之相反，敏感性持续下降。这种现象表明不平衡数据集会导致大部分样本被预测为占绝大多数样本的类别，再次验证了不平衡数据集问题确实影响了预测性能。另外，准确率从0.846逐渐升高到0.949，其变化趋势与敏感性恰恰相反，说明数据集的不平衡性越严重，准确率反而越高。因此，对于不平衡数据集的分类问题，准确率不是一个合理的测度。而随着负样本比例的提高，均衡准确率变化幅度较小。以上结果说明本发明利用均衡准确率指标来选择最优特征子集是合情合理的。

[0097] 为解决不平衡数据集问题，本发明将采用集成学习方法来识别细胞外基质蛋白。之前的研究结论认为集成分类器通常优于单个分类器，不仅能提高预测性能，而且能增加预测结果的可信度。

[0098] 本发明通过10-交叉验证方法评估训练数据集的预测性能。正负样本集分别随机地分为样本数量几乎相同的10组数据子集。在这生成的20组数据子集中，正负样本集的各一组数据子集用于训练，剩余的数据子集用于测试。每次用于训练的数据子集保证不同，上述过程循环重复10次。对于上述每一次过程，一种用于识别细胞外基质蛋白的集成学习方法如图5所示，其执行流程包括如下步骤。

[0099] 步骤一：训练数据集中负样本的个数大约为正样本的11倍，负样本训练集通过欠采样方法分为样本数量几乎相同的11组数据子集。每一组数据子集与正样本训练集构成训练子集。通过上述欠采样过程，可得到11个训练子集。

[0100] 步骤二：分别用步骤一得到的11个训练子集训练随机森林分类器，所获取的11个随机森林分类器组成集成分类器。测试样本集用于评估集成分类器的性能。基于集成分类器，测试样本的最终预测类别通过多数表决的方法获得。

[0101] 为验证集成学习方法在解决不平衡数据集问题方面的有效性，表1给出了有无集成学习方法的预测结果。如表1所示，在无集成学习方法时，特异性和准确率分别为0.956，0.989。由于数据集的不平衡性，敏感性仅仅为0.598。然而，集成学习方法具有较均衡的敏感性和特异性，分别为0.878，0.849。以上结果表明集成学习方法成功地解决了数据集不平衡的问题。

[0102] 表1有无集成学习方法的预测结果

[0103]

方法	敏感性	特异性	准确率	均衡准确率
无集成学习方法	0.598	0.989	0.956	0.793
有集成学习方法	0.878	0.849	0.851	0.864

[0104] 为更加客观地评估集成学习方法的预测能力，在独立测试样本集上，表2比较了本发明用于识别细胞外基质蛋白(Identify ECM Protein)的方法IECMP与先前的研究方法ECMPP、ECMPRED的预测结果。

[0105] 如表2所示，ECMPP得到了最低的敏感性和最高的特异性，这可能归因于数据集不平衡问题(410个正样本和4464个负样本)。尽管ECMPRED利用平衡的数据集(410个正样本和410个负样本)来训练，但由于没有充分利用训练数据集中负样本的信息，导致ECMPRED的特异性和均衡准确率达到了最低。另外，ECMPP和ECMPRED的敏感性和特异性差异都很大。而本发明的方法IECMP得到了较均衡的敏感性(0.765)和特异性(0.785)。对于均衡准确率指标，IECMP也远远好于ECMPP和ECMPRED。因此，对于细胞外基质蛋白识别问题，本发明的集成学习方法更优于先前的方法。

[0106] 表2本发明方法与现有方法在独立测试数据上的预测结果

[0107]

方法	敏感性	特异性	准确率	均衡准确率
ECMPP	0.294	0.985	0.712	0.640
ECMPRED	0.622	0.478	0.535	0.550
IECMP	0.765	0.785	0.777	0.775

[0108] 为方便用户使用本发明提出的方法来识别细胞外基质蛋白，我们开发了用于细胞外基质蛋白识别的网络服务器系统，用户键入网址“http://219.231.143.58/ch”可以免费访问。图6为ICEMP网络服务器的主页，图7为ICEMP网络服务器的预测结果页面。如图6所示，用户无需理解IECMP的执行过程，只需以FASTA格式输入待预测序列或输入待预测序列的UniprotKB ID，并点击提交按钮，此时ICEMP网络服务器会立即执行集成学习方法，在预测结果页面返回蛋白质的预测类别及其置信水平。下面详细说明使用ICEMP网络服务器的步骤。

[0109] 步骤一：键入网址“http://219.231.143.58/ch”，可以访问IECMP网络服务器的首页。在首页导航栏中点击“工具”链接，即可进入IECMP网络服务器页面。如需浏览IECMP网络服务器页面的使用指南，单击“帮助”链接；

[0110] 步骤二：以FASTA格式输入待预测序列或输入待预测序列的UniprotKB ID。单击“实例”链接，可以获取FASTA格式的具体形式。本服务器每次输入的蛋白质序列不应多于10个；

[0111] 步骤三：输入电子邮箱地址，点击提交按钮。ICEMP网络服务器会立即执行集成学习方法，获取预测结果后，系统立即发邮件通知，并在预测结果页面返回蛋白质的预测类别及其置信水平。

[0112] 步骤四：单击“下载”链接，转到数据集下载页面。用户可以免费下载本网络服务器用到的训练数据集和测试样本集。

[0113] 本发明涉及蛋白质功能属性识别领域，用于识别细胞外基质蛋白的集成学习方法，此集成学习方法全面综合了蛋白质的序列信息，包括序列组成、物理化学属性、进化信息和结构信息。并通过信息增益率-增量特征选择方法进一步提高预测性能，降低维数灾难。本发明提出的方法IECMP成功解决了细胞外基质蛋白识别过程中的数据集不平衡问题，得到了较均衡的敏感性和特异性。在独立测试样本集上，IECMP的预测结果优于先前的研究方法ECMPP和ECMPRED，验证了IECMP是一个有效的细胞外基质蛋白识别方法。该方法将辅助我们深入理解ECM蛋白相关的生物学过程机制，并为发现候选的药物靶点提供重要线索。为方便用户，基于集成学习方法，本发明开发了用于识别细胞外基质蛋白的网络服务器

IECMP。

[0114] 上述虽然结合附图对本发明的具体实施方式进行了描述，但并非对本发明保护范围的限制，所属领域技术人员应该明白，在本发明的技术方案的基础上，本领域技术人员不需要付出创造性劳动即可做出的各种修改或变形仍在本发明的保护范围以内。

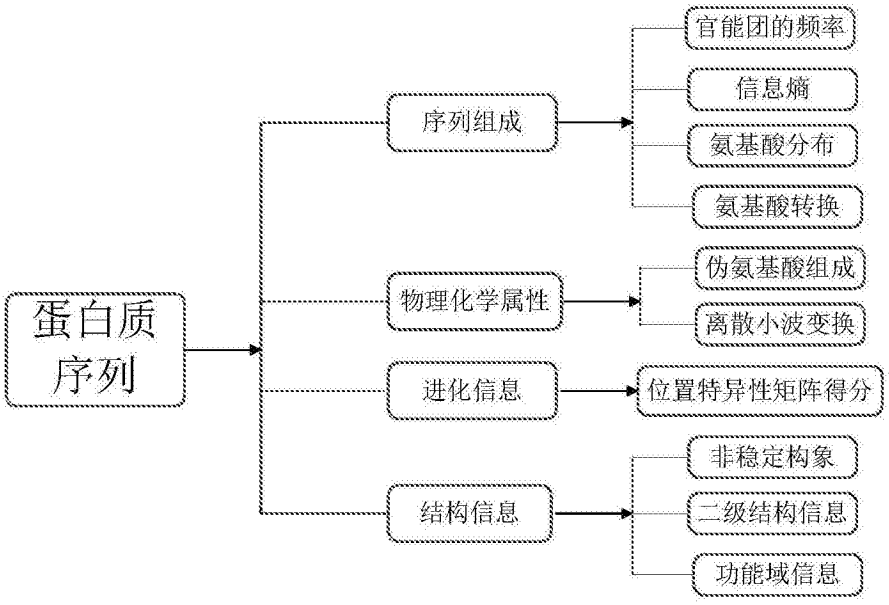

蛋白质序列
序列组成
物理化学属性
进化信息
结构信息
官能团的频率
信息熵
氨基酸分布
氨基酸转换
伪氨基酸组成
离散小波变换
位置特异性矩阵得分
非稳定构象
二级结构信息
功能域信息

图1

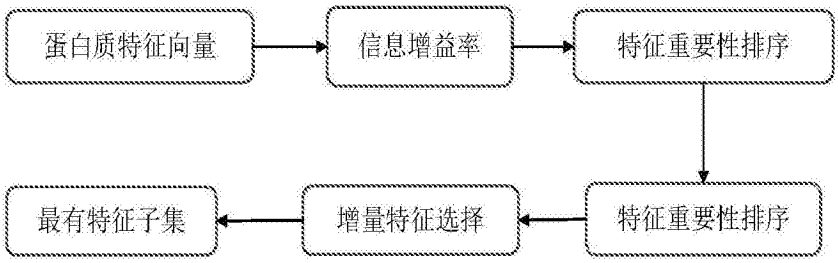

蛋白质特征向量
信息增益率
特征重要性排序
特征重要性排序
增量特征选择
最有特征子集

图2

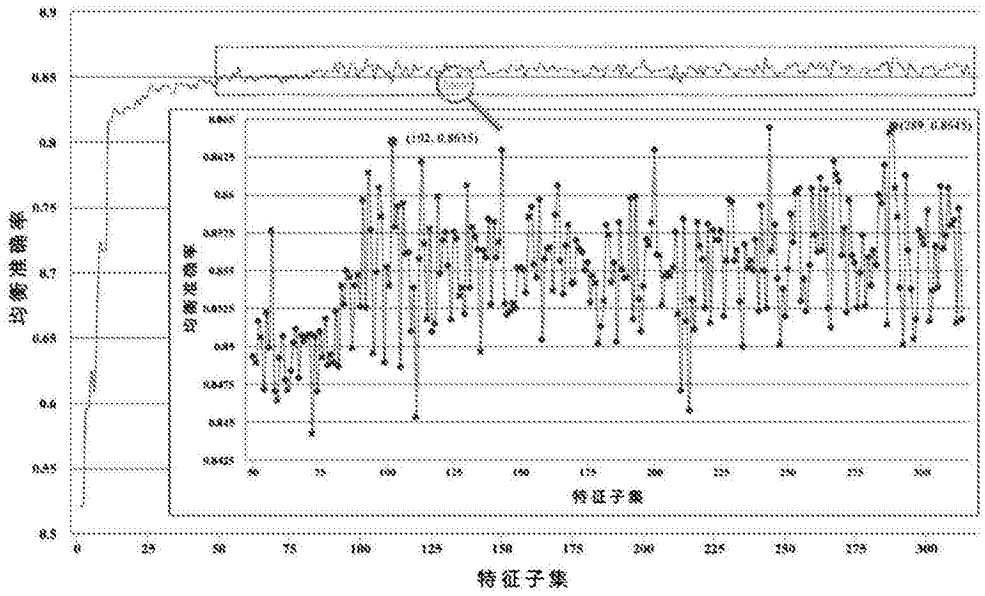


图3

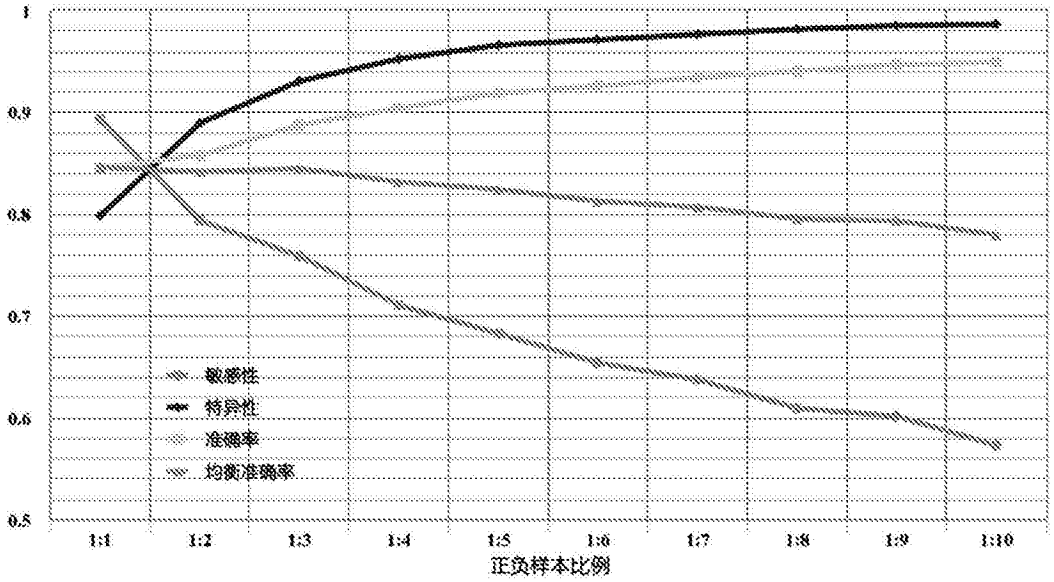


图4

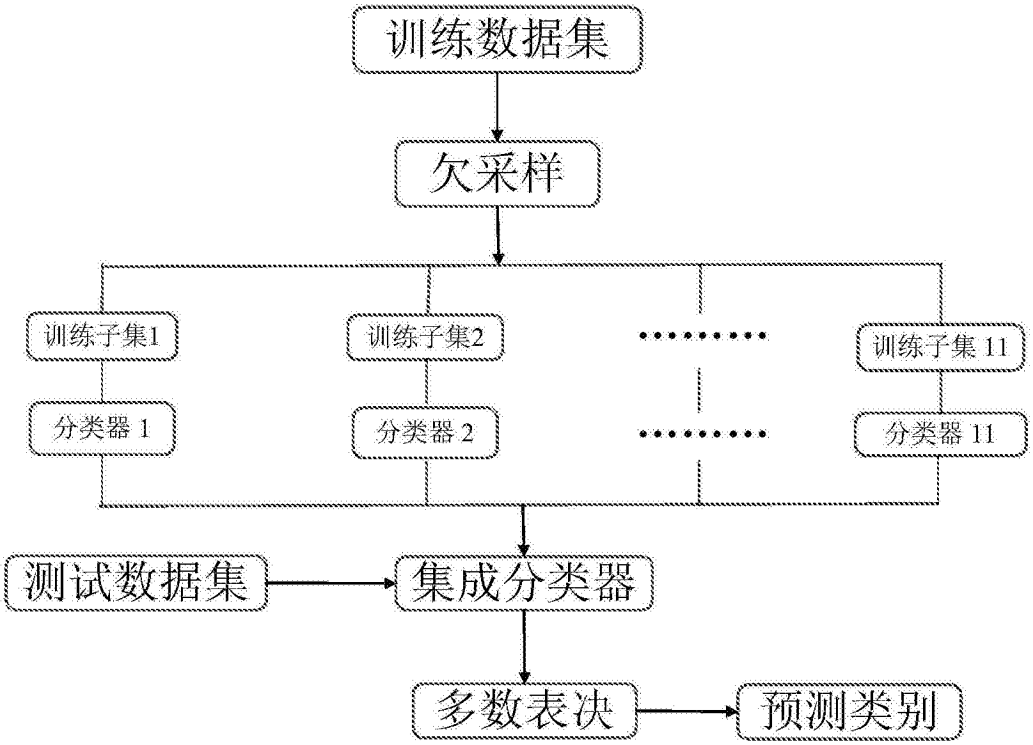

训练数据集
欠采样
训练子集1
训练子集2
训练子集 11
分类器 1
分类器 2
分类器 11
测试数据集
集成分类器
多数表决
预测类别

图5

图6

图7