

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 7 月 5 日 (05.07.2018)



(10) 国际公布号
WO 2018/120128 A1

- (51) 国际专利分类号:
G06F 19/18 (2011.01)
- (21) 国际申请号: PCT/CN2016/113754
- (22) 国际申请日: 2016 年 12 月 30 日 (30.12.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。
- (72) 发明人: 张慧玲 (ZHANG, Huiling); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 魏彦杰 (WEI, Yanjie);

中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 郭宁 (GUO, Ning); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 贝振东 (BEI, Zhendong); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 朱昱寰 (ZHU, Yuhuan); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。

- (74) 代理人: 深圳中一专利商标事务所 (SHENZHEN ZHONGYI PATENT AND TRADEMARK OFFICE); 中国广东省深圳市福田区深南中路 1014 号老特区报社四楼 (5 号信箱), Guangdong 518028 (CN)。
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: METHOD AND DEVICE FOR PREDICTING INTERACTION RELATIONSHIP BETWEEN MEMBRANE PROTEIN RESIDUES

(54) 发明名称: 一种膜蛋白残基的作用关系的预测方法和装置

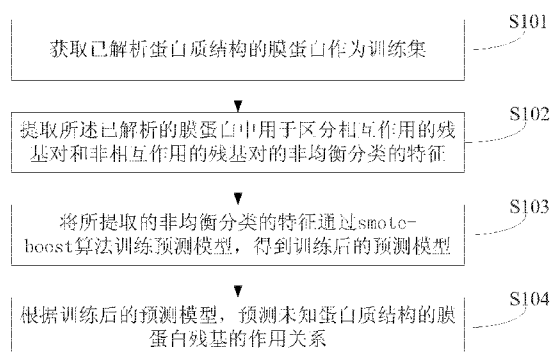


图 1

- S101 Obtaining a membrane protein where the protein structure of same has been analyzed to be a training set
- S102 Extracting the features of the unbalanced classification for distinguishing interacting residue pairs from non-interacting residue pairs in the resolved membrane protein
- S103 Training a predictive model with the extracted features of the unbalanced classification by means of a SMOTEBoost algorithm, and obtaining the trained predictive model
- S104 Predicting the interaction relationship between membrane protein residues where the protein structures are unknown according to the trained predictive model

(57) Abstract: Disclosed is a method for predicting the interaction relationship between membrane protein residues, comprising: obtaining a membrane protein, where the protein structure of same has been analyzed to be a training set; extracting the features of the unbalanced classification for distinguishing interacting residue pairs from non-interacting residue pairs in the membrane protein where the protein structure has been resolved; training a predictive model with the extracted features of the unbalanced classification by means of a SMOTE-Boost algorithm; and predicting the interaction relationship between membrane protein residues where the protein structures are unknown according to the trained predictive model. Due to the use of the features of the unbalanced classification to train the predictive model, the trained predictive model can avoid the loss of useful information, and can help to improve the accuracy and coverage of the prediction.

(57) 摘要: 一种膜蛋白残基的作用关系的预测方法包括: 获取已解析蛋白质结构的膜蛋白作为训练集; 提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征; 将所提取的非均衡分类的特征通过smote-boost算法训练预测模型; 根据训练后的预测模型, 预测未知蛋白质结构的膜蛋白残基的作用关系。由于使用非均衡分类的特征进行预测模型的训练, 从而使得训练后的预测模型能够避免有用信息的流失, 有利于提高预测的精准度和覆盖度。

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

说明书

发明名称：一种膜蛋白残基的作用关系的预测方法和装置

技术领域

[0001] 本发明属于数据挖掘、机器学习和计算机生物学的交叉领域，尤其涉及一种膜蛋白残基作用关系的预测方法和装置。

背景技术

[0002] 在目前已知的药物靶点中，膜蛋白约占60%。由于膜蛋白结构的实验解析难度较大，在蛋白质数据库（Protein Data Bank-PDB）中，超过9万个的已知蛋白质结构里，已知的膜蛋白结构仅占已知的蛋白质结构的1%。

[0003] 现有的解析蛋白质三维结构的生物学实验方法主要包括X-RAY和NMR法。这些生物学实验方法不仅操作过程较为复杂，耗时，而且实验花费的成本也较高。正是由于实验解析法的这些不足，使得计算机计算方法的发展成为必然。目前用于蛋白质三维结构预测的计算方法主要有同源模建法、折叠识别法和从头预测法。并且通常从均衡分类的角度，将相互作用的残基对或非相互作用的残基对按照1: 1的比例训练模型。其中，残基是指由20种不同的氨基酸连接形成的多聚体，在形成蛋白质后，这些氨基酸之间的氨基和羧基脱水成键，氨基酸由于其部分基团参与了肽键的形成，剩余的结构部分称为氨基酸残基。所谓残基作用关系是指那些在蛋白质的一级序列中不相邻而在三级结构中邻近的残基对。

[0004] 由于相互作用的残基对与非相互作用的残基对的比例一般会远远大于1: 1，从而使得现有的预测方法会导致大量有用的信息流失，影响预测的准确度和覆盖度。

技术问题

[0005] 本发明的目的在于提供一种膜蛋白残基的作用关系的预测方法，以解决现有技术中的预测方法会导致大量有用的信息流失，影响预测的准确度和覆盖度的问题。

问题的解决方案

技术解决方案

- [0006] 第一方面，本发明实施例提供了一种膜蛋白残基的作用关系的预测方法，所述方法包括：
- [0007] 获取已解析蛋白质结构的膜蛋白作为训练集；
- [0008] 提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；
- [0009] 将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；
- [0010] 根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。
- [0011] 结合第一方面，在第一方面的第一种可能实现方式中，所述提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征步骤中，所述非均衡分类的特征包括：位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。
- [0012] 结合第一方面的第一种可能实现方式，在第一方面的第二种可能实现方式中，所述位置特异性得分矩阵PSSM中的每个残基由一个20维的向量表示，所述位置特异性得分矩阵PSSM特征包括：
- [0013] 以残基对 (i,j) 中的残基i和残基j分别为中心取一个大小为a的滑动容器，每个残基对得到 $40a$ 个位置特异性得分矩阵PSSM特征；
- [0014] 以残基对 (i,j) 的中间位置 $(i+j)/2$ 为中心取一个大小为b的滑动窗口，获得 $20*b$ 个位置特异性得分矩阵PSSM特征。
- [0015] 结合第一方面的第一种可能实现方式，在第一方面的第三种可能实现方式中，一个残基作用对包括两个氨基酸，所述残基类型特征包括由酸性氨基酸、碱性氨基酸、极性氨基酸、非极性氨基酸中的任意两种所产生的10种组合。
- [0016] 结合第一方面，在第一方面的第四种可能实现方式中，所述相互作用的残基对为位于膜蛋白的 α 螺旋上的CB-CB原子距离小于8埃的残基对。
- [0017] 第二方面，本发明实施例提供了一种膜蛋白残基的作用关系的预测装置，所述装置包括：

- [0018] 训练集获取单元，用于获取已解析蛋白质结构的膜蛋白作为训练集；
- [0019] 特征提取单元，用于提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；
- [0020] 训练单元，用于将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；
- [0021] 预测单元，用于根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。
- [0022] 结合第二方面，在第二方面的第一种可能实现方式中，所述特征提取单元中，所述非均衡分类的特征包括：位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。
- [0023] 结合第二方面的第一种可能实现方式，在第二方面的第二种可能实现方式中，所述位置特异性得分矩阵PSSM中的每个残基由一个20维的向量表示，所述位置特异性得分矩阵PSSM特征包括：
- [0024] 以残基对 (i,j) 中的残基 i 和残基 j 分别为中心取一个大小为 a 的滑动容器，每个残基对得到 $40a$ 个位置特异性得分矩阵PSSM特征；
- [0025] 以残基对 (i,j) 的中间位置 $(i+j)/2$ 为中心取一个大小为 b 的滑动窗口，获得 $20*b$ 个位置特异性得分矩阵PSSM特征。
- [0026] 结合第二方面的第一种可能实现方式，在第二方面的第三种可能实现方式中，一个残基作用对包括两个氨基酸，所述残基类型特征包括由酸性氨基酸、碱性氨基酸、极性氨基酸、非极性氨基酸中的任意两种所产生的10种组合。
- [0027] 结合第二方面，在第二方面的第四种可能实现方式中，所述相互作用的残基对为位于膜蛋白的 α 螺旋上的CB-CB原子距离小于8埃的残基对。

发明的有益效果

有益效果

- [0028] 在本发明中，获取已解析的蛋白质结构的膜蛋白作为训练集，提取所述已解析的蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征，将提取的特征通过smote-boost算法训练预测模型，得到训练

后的预测模型，并根据所述训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。由于使用非均衡分类的特征进行预测模型的训练，从而使得训练后的预测模型能够避免有用信息的流失，有利于提高预测的精准度和覆盖度。

对附图的简要说明

附图说明

[0029] 图1是本发明实施例提供的膜蛋白残基的作用关系的预测方法的实现流程图；

[0030] 图2是本发明实施例提供的膜蛋白残基的作用关系的预测装置的结构示意图。

发明实施例

本发明的实施方式

[0031] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0032] 本发明实施例的目的在于提供一种膜蛋白残基的作用关系的预测方法，以解决现有技术中对于未知结构的膜蛋白残基的作用关系的预测过程中，一般从均衡分类的角度将相互作用的残基对或非相互作用的残基对按照1: 1的比例训练模型，而实际上，相互作用或非相互作用的残基对比例远远大于1: 1，按照均衡对等的比例训练模型会造成大量有用信息的流失，从而会导致预测的膜蛋白残基的作用关系的精准度和覆盖度不高的问题。下面结合附图对本发明作进一步的说明。

[0033] 图1示出了本发明第一实施例提供的膜蛋白残基的作用关系的预测方法的实现流程，详述如下：

[0034] 在步骤S101中，获取已解析蛋白质结构的膜蛋白作为训练集。

[0035] 具体的，所述已解析蛋白质结构的膜蛋白，应当已确定的膜蛋白残基的作用关系。优选的一种实施方式，可以使用PDBTM(英文全称为：protein data bank of transmembrane proteins，中文全称为：跨膜蛋白的蛋白质数据库)中2012年2月以前解析的膜蛋白作为训练集。

[0036] 当然，上述膜蛋白数据库的训练集的选取只是其中一种优选的实施方式，随着

解析和识别技术的发展，越来越多的膜蛋白结构被解析，能够得到确定的膜蛋白残基的作用关系，因而所述训练集中的样本数据也会越来越丰富，因而也会更加有利于提高预测模型的训练的准确度。

[0037] 在步骤S102中，提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；

[0038] 具体的，本发明实施例中所述用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征，可以包括位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。

[0039] 其中，所述位置特异性得分矩阵PSSM(英文全称为：Position-Specific Scoring Matrix)特征，可以通过运行PSI-BLAST(英文全称为：Position-Specific Iterative Basic Local Alignment Search Tool，中文全称为：位置特异性迭代搜索算法)的方式获取。其中，运行PSI-BLAST时可以采用的数据库是UNIREF90数据库，运行时的迭代次数可以为2，E-value截断值为 $1e-10$ （表示为 1×10^{-10} 次方）。

[0040] 在本发明实施例中，所述位置特异性得分矩阵PSSM中的每个残基都由一个20维的向量表示，表示20种氨基酸在PSSM相应位置出现的频率。特征提取时，位置特异性得分矩阵PSSM特征分为两类，分别为：

[0041] 以残基对 (i, j) 中的残基 i 和残基 j 分别为中心取一个大小为 a 的滑动容器，每个残基对得到 $40a$ 个位置特异性得分矩阵PSSM特征；

[0042] 以残基对 (i, j) 的中间位置 $(i+j)/2$ 为中心取一个大小为 b 的滑动窗口，获得 $20 \times b$ 个位置特异性得分矩阵PSSM特征。

[0043] 比如，具体的一种实施方式中，可以为：

[0044] 第一类是以残基对 (i, j) 中的残基 i 和残基 j 分别为中心取一个大小为7的滑动窗口，即对每个残基对可得到 $2 \times 7 \times 20 = 280$ 个位置特异性得分矩阵PSSM特征；

[0045] 第二类是以残基对 (i, j) 的中间位置 $(i+j)/2$ 为中心取一个大小为3的滑动窗口，即可获得 $3 \times 20 = 60$ 个位置特异性得分矩阵PSSM特征。

[0046] 两类位置特异性得分矩阵PSSM特征的总数为 $280 + 60 = 340$ 个。

[0047] 所述残基在 α 螺旋中相对距离特征具体为：假设 p 为残基对中的一个残基在长

度为1的螺旋上的相对位置，那么残基在 α 螺旋中相对距离特征就定义为 p/l ，对于每个残基对中包括两个残基，可以分别提取残基所对应的残基在 α 螺旋中相对距离特征，一共包括2个残基在 α 螺旋中相对距离特征。

[0048] 所述序列间隔特征可以根据残基对在一级序列中的位置进行划分。比如，一种具体的间隔划分方式可以划分为以下多个区间：

[0049] <25 、 $25-50$ 、 $50-75$ 、 $75-100$ 、 $100-125$ 、 $125-150$ 、 $150-175$ 、 $175-200$ 和 >200 这九个区间。

[0050] 可将使用相应的序列间隔特征码000000000置0或置1（0表示不在该区间，反之为1）用于表述序列间隔特征。对于每个残基对而言，按照上述区间划分方式，可以对应9个序列间隔特征中的一个。

[0051] 对于所述残基类型特征，考虑到组成蛋白质的氨基酸共20种，根据氨基酸R基的极性性质可分为酸性氨基酸（谷氨酸及天冬氨酸）、碱性氨基酸（赖氨酸、精氨酸及组氨酸）和中性氨基酸，其中中性氨基酸又可分为极性氨基酸（甘氨酸、丝氨酸、半胱氨酸、苏氨酸、酪氨酸、天冬酰胺及谷氨酰胺）和非极性氨基酸（丙氨酸、亮氨酸、异亮氨酸、苯丙氨酸、甲硫氨酸、色氨酸、缬氨酸及脯氨酸）。根据这4种不同的氨基酸类型（酸性氨基酸、碱性氨基酸、极性氨基酸和非极性氨基酸），一个残基作用对（对应两个氨基酸）可以产生10种不同的组合，可以二进制码0000000000分别置0或置1来代表不同的组合类型。可以包括10个残基类型特征。

[0052] 所述 α 螺旋个数特征可以根据膜蛋白所包含的 α 螺旋个数进行区间划分。比如，可以划分为2-4、5-7、8-10、以及大于10这4个区间。通过二进制向量0000置0或置1来表示该 α 螺旋个数特征（0表示不在该区间，反之为1）。该类特征对某一膜蛋白中所有残基对具有一致性。每个残基对特征向量包含4个该类特征。

[0053] 所述序列长度特征，可以根据膜蛋白所一级序列的长度可分为 <100 ， $100-400$ ， $400-800$ ， >800 这4个区间，以二进制向量0000置0或置1来表示该特征（0表示不在该区间，反之为1）。这类特征对同一个膜蛋白中的所有残基对均一致。每个残基对特征向量包含4个该类特征。

[0054] 综上所述，本发明可以使用340个位置特异性得分矩阵PSSM特征，2个 α 螺旋

中相对距离特征，9个序列间隔特征以及10个残基类型特征，4个 α 螺旋个数特征，4个序列长度特征，共计369个特征。

[0055] 另外，本发明实施例中所述相互作用的残基对和非相互作用的残基对的比例，可以为1比50至1比80，优选的一种实施方式可以设置为1比67。

[0056] 具体的，蛋白质残基作用对的定义有多种，例如基于原子的范德华距离的定义，基于CA-CA原子距离的定义以及基于CB-CB原子距离的定义。本发明关于残基作用对的定义将沿用被广泛采用的定义：将位于膜蛋白的 α 螺旋上的CB-CB原子距离小于 $8\text{\AA}$ （埃）

的残基对定义为相互作用的残基对。CA、CB是gromacs里面的原子类型，gromacs分子动力学软件。

[0057] 在步骤S103中，将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；

[0058] 在提到到所述非均衡分类的特征后，可以将所述特征代入到预测模型中进行训练。所述预测模型可以为向量机训练模型等。

[0059] 所述训练算法smote-boost，是将smote技术和boost技术结合的新型训练方法，其中：boost方法在每次迭代中，增加没有正确分类样本的权值，减少正确分类样本的权值，更加关注于分类错误的样本。因为少数样本更容易被错误分类，所以该方法能够改进对少数类的预测性能。SMOTE(英文全称为synthetic minority over-sampling rechnique)技术是非均衡数据集学习的一种新办法，通过对少数样本的人工合成提高少数类样本的比例，降低数据的过度偏斜。SMOTE技术与BOOST技术相结合，可以有效避免由于赋予少数样本更大权值可能产生的过度拟合。

[0060] 在步骤S104中，根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。

[0061] 本发明通过获取已解析的蛋白质结构的膜蛋白作为训练集，提取所述已解析的蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征，将提取的特征通过smote-boost算法训练预测模型，得到训练后的预测模型，并根据所述训练后的预测模型，预测未知蛋白质结构的膜蛋白残

基的作用关系。由于使用非均衡分类的特征进行预测模型的训练，从而使得训练后的预测模型能够避免有用信息的流失，有利于提高预测的精准度和覆盖度。

[0062] 图2示出了本发明实施例提供的一种膜蛋白残基的作用关系的预测装置的结构示意图，详述如下：

[0063] 本发明实施例所述膜蛋白残基的作用关系的预测装置，包括：

[0064] 训练集获取单元201，用于获取已解析蛋白质结构的膜蛋白作为训练集；

[0065] 特征提取单元202，用于提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；

[0066] 训练单元203，用于将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；

[0067] 预测单元204，用于根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。

[0068] 优选的，所述特征提取单元中，所述非均衡分类的特征包括：位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。

[0069] 优选的，所述位置特异性得分矩阵PSSM中的每个残基由一个20维的向量表示，所述位置特异性得分矩阵PSSM特征包括：

[0070] 以残基对 (i,j) 中的残基 i 和残基 j 分别为中心取一个大小为 a 的滑动容器，每个残基对得到 $40a$ 个位置特异性得分矩阵PSSM特征；

[0071] 以残基对 (i,j) 的中间位置 $(i+j)/2$ 为中心取一个大小为 b 的滑动窗口，获得 $20*b$ 个位置特异性得分矩阵PSSM特征。

[0072] 优选的，一个残基作用对包括两个氨基酸，所述残基类型特征包括由酸性氨基酸、碱性氨基酸、极性氨基酸、非极性氨基酸中的任意两种所产生的10种组合。

[0073] 优选的，所述相互作用的残基对为位于膜蛋白的 α 螺旋上的CB-CB原子距离小于8埃的残基对。

[0074] 图2所述膜蛋白残基的作用关系的预测装置，与实施例一所述膜蛋白残基的作

用关系的预测方法对应，在此不作重复赘述。

[0075] 在本发明所提供的几个实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0076] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0077] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0078] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施例所述方法的全部或部分。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM, Read-Only Memory)、随机存取存储器（RAM, Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

[0079] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

权利要求书

- [权利要求 1] 一种膜蛋白残基的作用关系的预测方法，其特征在于，所述方法包括：
- 获取已解析蛋白质结构的膜蛋白作为训练集；
- 提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；
- 将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；
- 根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。
- [权利要求 2] 根据权利要求1所述方法，其特征在于，所述提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征步骤中，所述非均衡分类的特征包括：位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。
- [权利要求 3] 根据权利要求2所述方法，其特征在于，所述位置特异性得分矩阵PSSM中的每个残基由一个20维的向量表示，所述位置特异性得分矩阵PSSM特征包括：
- 以残基对 (i,j) 中的残基i和残基j分别为中心取一个大小为a的滑动容器，每个残基对得到 $40a$ 个位置特异性得分矩阵PSSM特征；
- 以残基对 (i,j) 的中间位置 $(i+j)/2$ 为中心取一个大小为b的滑动窗口，获得 $20*b$ 个位置特异性得分矩阵PSSM特征。
- [权利要求 4] 根据权利要求2所述方法，其特征在于，一个残基作用对包括两个氨基酸，所述残基类型特征包括由酸性氨基酸、碱性氨基酸、极性氨基酸、非极性氨基酸中的任意两种所产生的10种组合。
- [权利要求 5] 根据权利要求1所述方法，其特征在于，所述相互作用的残基对为位于膜蛋白的 α 螺旋上的CB-CB原子距离小于 8埃的残基对。

- [权利要求 6] 一种膜蛋白残基的作用关系的预测装置，其特征在于，所述装置包括：
- 训练集获取单元，用于获取已解析蛋白质结构的膜蛋白作为训练集；
- 特征提取单元，用于提取所述已解析蛋白质结构的膜蛋白中用于区分相互作用的残基对和非相互作用的残基对的非均衡分类的特征；
- 训练单元，用于将所提取的非均衡分类的特征通过smote-boost算法训练预测模型，得到训练后的预测模型；
- 预测单元，用于根据训练后的预测模型，预测未知蛋白质结构的膜蛋白残基的作用关系。
- [权利要求 7] 根据权利要求6所述装置，其特征在于，所述特征提取单元中，所述非均衡分类的特征包括：位置特异性得分矩阵PSSM特征、残基在 α 螺旋中相对距离特征、序列间隔特征、残基类型特征、 α 螺旋个数特征、序列长度特征中的一种或者多种。
- [权利要求 8] 根据权利要求7所述装置，其特征在于，所述位置特异性得分矩阵PSSM中的每个残基由一个20维的向量表示，所述位置特异性得分矩阵PSSM特征包括：
- 以残基对(i,j)中的残基i和残基j分别为中心取一个大小为a的滑动容器，每个残基对得到40a个位置特异性得分矩阵PSSM特征；
- 以残基对(i,j)的中间位置 $(i+j)/2$ 为中心取一个大小为b的滑动窗口，获得20*b个位置特异性得分矩阵PSSM特征。
- [权利要求 9] 根据权利要求7所述装置，其特征在于，一个残基作用对包括两个氨基酸，所述残基类型特征包括由酸性氨基酸、碱性氨基酸、极性氨基酸、非极性氨基酸中的任意两种所产生的10种组合。
- [权利要求 10] 根据权利要求6所述装置，其特征在于，所述相互作用的残基对为位于膜蛋白的 α 螺旋上的CB-CB原子距离小于 8埃的残基对。

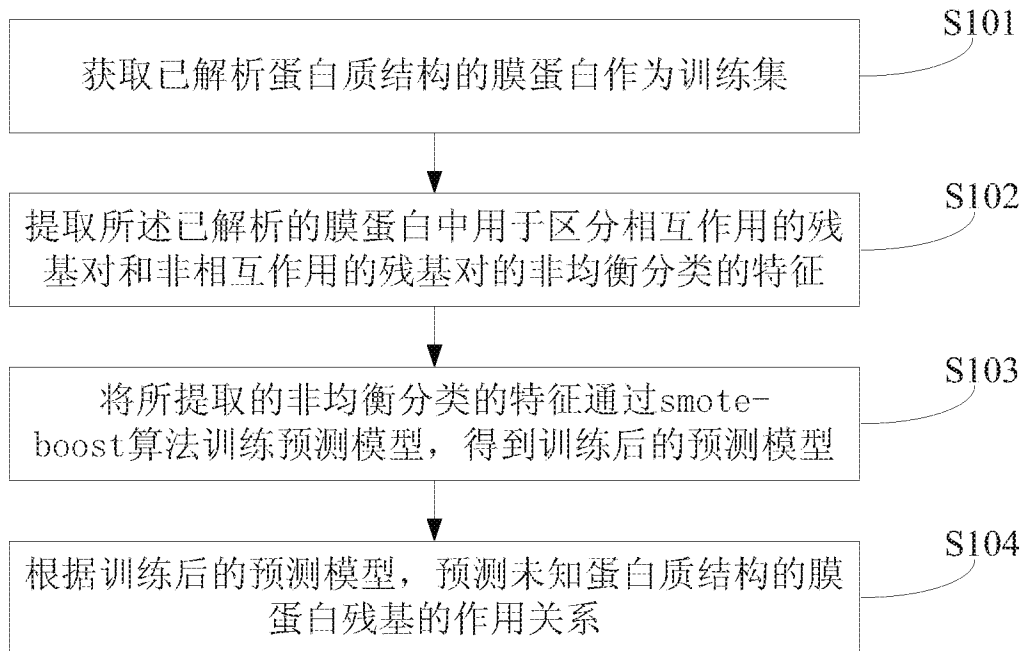


图 1

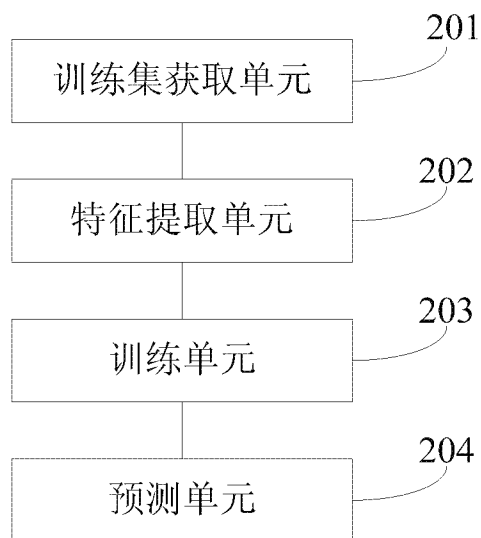


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/113754

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/18 (2011.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CPRSABS; CJFD; DWPI; SIPOABS; CNTXT; EPTXT; USTXT; WOTXT; CNKI; GOOGLE; PubMed; ISI Web of Knowledge; Patentics: 中国科学院深圳先进技术研究, 朱昱寰, 郭宁, 魏彦杰, 贝振东, 张慧玲, 膜蛋白, 残基, 作用, 非均衡分类, 算法, 模型, smote-boost, membrane protein, residue?, interact+, non-equilibrium classification, algorithm, model, imbalance, SMOTE, Random Forest, SMOTEBoost

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 106650309 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 10 May 2017 (10.05.2017), claims 1-10	1-10
Y	CN 104615910 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 13 May 2015 (13.05.2015), claims 1-10	1-10
Y	ARUN, K.M.N. et al., "On the Classification of Imbalanced Datasets", International Journal of Computer Applications, 44(08), 30 April 2012 (30.04.2012), pages 1-7	1-10
Y	FATTAHI, S. et al., "New Approach for Imbalanced Biological Dataset Classification", Journal of Theoretical and Applied Information Technology, 72(01), 10 February 2015 (10.02.2015), pages 40-51	1-10
A	CN 104252581 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 31 December 2014 (31.12.2014), entire document	1-10

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 15 August 2017	Date of mailing of the international search report 08 September 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer CHEN, Ying Telephone No. (86-10) 62413889

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/113754

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104504299 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 08 April 2015 (08.04.2015), entire document	1-10

International application No.
PCT/CN2016/113754

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类 G06F 19/18 (2011.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																				
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS; CPRSABS; CJFD; DWPI; SIPOABS; CNTXT; EPTXT; USTXT; WOTXT; CNKI; GOOGLE; PubMed; ISI Web of Knowledge; Patents: 中国科学院深圳先进技术研究院, 朱昱寰, 郭宁, 魏彦杰, 贝振东, 张慧玲, 膜蛋白, 残基, 作用, 非均衡分类, 算法, 模型, smote-boost, membrane protein, residue?, interact+, non-equilibrium classification, algorithm, model, imbalance, SMOTE, Random Forest, SMOTEBoost																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>E</td> <td>CN 106650309 A (中国科学院深圳先进技术研究院) 2017年 5月 10日 (2017 - 05 - 10) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>CN 104615910 A (中国科学院深圳先进技术研究院) 2015年 5月 13日 (2015 - 05 - 13) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>ARUN, K. M. N. 等. "On the Classification of Imbalanced Datasets" International Journal of Computer Applications, 第44卷, 第08期, 2012年 4月 30日 (2012 - 04 - 30), 第1-7页</td> <td>1-10</td> </tr> <tr> <td>Y</td> <td>FATTAHI, S. 等. "NEW APPROACH FOR IMBALANCED BIOLOGICAL DATASET CLASSIFICATION" Journal of Theoretical and Applied Information Technology, 第72卷, 第01期, 2015年 2月 10日 (2015 - 02 - 10), 第40-51页</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 104252581 A (中国科学院深圳先进技术研究院) 2014年 12月 31日 (2014 - 12 - 31) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	E	CN 106650309 A (中国科学院深圳先进技术研究院) 2017年 5月 10日 (2017 - 05 - 10) 权利要求1-10	1-10	Y	CN 104615910 A (中国科学院深圳先进技术研究院) 2015年 5月 13日 (2015 - 05 - 13) 权利要求1-10	1-10	Y	ARUN, K. M. N. 等. "On the Classification of Imbalanced Datasets" International Journal of Computer Applications, 第44卷, 第08期, 2012年 4月 30日 (2012 - 04 - 30), 第1-7页	1-10	Y	FATTAHI, S. 等. "NEW APPROACH FOR IMBALANCED BIOLOGICAL DATASET CLASSIFICATION" Journal of Theoretical and Applied Information Technology, 第72卷, 第01期, 2015年 2月 10日 (2015 - 02 - 10), 第40-51页	1-10	A	CN 104252581 A (中国科学院深圳先进技术研究院) 2014年 12月 31日 (2014 - 12 - 31) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
E	CN 106650309 A (中国科学院深圳先进技术研究院) 2017年 5月 10日 (2017 - 05 - 10) 权利要求1-10	1-10																		
Y	CN 104615910 A (中国科学院深圳先进技术研究院) 2015年 5月 13日 (2015 - 05 - 13) 权利要求1-10	1-10																		
Y	ARUN, K. M. N. 等. "On the Classification of Imbalanced Datasets" International Journal of Computer Applications, 第44卷, 第08期, 2012年 4月 30日 (2012 - 04 - 30), 第1-7页	1-10																		
Y	FATTAHI, S. 等. "NEW APPROACH FOR IMBALANCED BIOLOGICAL DATASET CLASSIFICATION" Journal of Theoretical and Applied Information Technology, 第72卷, 第01期, 2015年 2月 10日 (2015 - 02 - 10), 第40-51页	1-10																		
A	CN 104252581 A (中国科学院深圳先进技术研究院) 2014年 12月 31日 (2014 - 12 - 31) 全文	1-10																		
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: "A" 认为不特别相关的表示了现有技术一般状态的文件 "E" 在国际申请日的当天或之后公布的在先申请或专利 "L" 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) "O" 涉及口头公开、使用、展览或其他方式公开的文件 "P" 公布日先于国际申请日但迟于所要求的优先权日的文件 "T" 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 "X" 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 "Y" 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 "&" 同族专利的文件																				
国际检索实际完成的日期		国际检索报告邮寄日期																		
2017年 8月 15日		2017年 9月 8日																		
ISA/CN的名称和邮寄地址		受权官员																		
中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		陈莹																		
		电话号码 (86-10) 62413889																		

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 104504299 A (中国科学院深圳先进技术研究院) 2015年 4月 8日 (2015 - 04 - 08) 全文	1-10

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/113754

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106650309	A	2017年 5月 10日	无	
CN	104615910	A	2015年 5月 13日	无	
CN	104252581	A	2014年 12月 31日	无	
CN	104504299	A	2015年 4月 8日	无	