Neurocomputing 228 (2017) 256–262

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

DrugRPE：用于药物-靶点相互作用预测的随机投影集成方法

Jun Zhang[a], Muchun Zhu[a], Peng Chen[a,], Bing Wang[b,]

中国安徽省合肥市安徽大学电气工程与自动化学院健康科学研究所，邮编：230601

[b] 同济大学电子与信息工程学院智能感知网络高级研究所，上海市嵌入式系统与服务计算重点实验室，中国上海 201804

文章信息

2015 年 MSC:
00-01
99-00
关键词：随机投影 药物 - 靶点相互作用 REP Tree 集成系统

摘 要

药物与靶点的相互作用在药物研发中至关重要。由于通过体外实验确定药物与靶点的相互作用既昂贵又耗时，因此计算方法成为确定相互作用的一种补充手段。为解决这一问题，提出了一种随机投影集成方法。首先，利用“PaDEL-Descriptor”软件对药物化合物进行特征描述符编码。其次，对靶蛋白用氨基酸的理化性质进行编码，从 AAindex1 数据库的 544 种性质中提取出 34 种相对独立的理化性质。对药物 - 靶点对的向量进行不同维度的随机投影，可以将原始空间投影到一个降维空间，从而得到一个固定维度的变换向量。多个随机投影构建了一个集成的 REPTree 系统。实验结果表明，在常用的药物 - 靶点基准数据集上，我们的方法显著优于其他最先进的药物 - 靶点预测器，并且运行速度更快。

1. 简介

药物与靶点相互作用是指确定一对药物和靶点能否相互作用。这是针对特定疾病进行药物研发的关键所在[1]。在合成候选药物之前[2,3]，需要克服几个难题。第一个难题是如何找出药物对不同人群的影响[4-6]，第二个难题是追踪并阐明药物在人体内沿生物相互作用途径的作用[7]。此外，由于药物研发成本高昂且耗时，每年获批的新药数量又相当少，因此计算方法成为药物研发的有力补充。计算方法可用于在候选药物获批之前确定其敏感性和毒性[2,3]，并且能在很大程度上节省时间和资金。

许多研究工作开发出了不同的计算方法来分析和识别药物与靶点的相互作用。这些方法可以分为多种类别：对接模拟[8,9]、文献文本挖掘[10]、结合化学结构、基因组序列和三维结构信息的方法[11,12]、基于核函数的方法[13]以及其他方法[14]。最常用的机器学习方法已被广泛应用于研究药物与靶点相互作用的问题。一些研究侧重于 HIV 蛋白酶切割位点预测[15]、G 蛋白偶联受体（GPCR）类型的识别[16]、蛋白质亚细胞定位预测[17,18]、膜蛋白类型的预测[19]，以及一系列相关的网络服务器预测工具，这些内容在最近的一篇综述中有所总结[20]。

机器学习方法在蛋白质相互作用领域中应用广泛[21-26]。在此，我们提出一种基于 REPTree 算法[27]的随机投影集成方法来预测药物-靶点相互作用，通过随机投影[28,29]将原始数据投影到一个较小的空间。为了将输入编码到分类器集成中，药物化合物通过“PaDEL-Descriptor”软件用特征描述符进行编码，而靶点蛋白质则用氨基酸的理化性质进行编码。从 AAindex1 数据库中的 544 种性质中，提取了 34 种相对独立的理化性质。对药物-靶点对的向量进行不同维度的随机投影，可以将原始空间映射到一个降维空间，从而得到一个固定维度的变换向量。本研究中，药物的蛋白质靶点与参考文献[11,12]相同，分为酶、离子通道、G 蛋白偶联受体（GPCRs）和核受体。多个随机投影构建了一个集成的 REPTree 系统。实验结果表明，在常用的药物-靶点基准数据集上，我们的方法显著优于其他最先进的药物-靶点预测器，并且运行速度更快。

Corresponding author.
E-mail addresses: junzhang@ahu.edu.cn (J. Zhang), bigeagle@mail.ustc.edu.cn (P. Chen), wangbing@ustc.edu (B. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.10.039
Received 29 December 2015; Received in revised form 4 April 2016; Accepted 24 October 2016 Available online 01 November 2016
0925-2312/

2. 方法

2.1. 表示目标蛋白质的特征向量

为了对目标蛋白质进行编码，使用了 AAindex1 数据库，该数据库包含 544 种氨基酸属性[30]。其中大多数属性都是相关的，因此，与我们之前的工作[31]一样，提取了相关系数（CC）小于 0.5 的不相关属性，就像 AAindex1 本身那样，它只呈现相关系数为 0.8 的相关属性。计算每两个属性的相关系数，并统计相关属性的数量。按相关数量降序排列，得到一个属性列表。对于排名首位的属性，我们移除所有与之相关的后续属性。依次类推，从列表中移除与前一个属性相关的每个属性。最终保留了 34 种属性，其中每两个属性的相关系数均小于 0.5 [31]。

对于第 i 个目标蛋白质链，本研究考虑了整个链中的所有残基。为了研究蛋白质残基在理化性质方面的进化，采用了一种结合氨基酸性质和序列特征的编码方案来表示残基。通过使用默认参数的 PSI-Blast [32] 为一个残基创建的序列特征，然后将其与每个氨基酸性质相乘，即一个氨基酸的性质乘以该氨基酸的序列特征得分。也就是说，对于残基 k 的特征 SP^k 和一个氨基酸性质尺度 Aap，它们都是具有 1×20 维度且氨基酸顺序相同的向量。此后，对于残基 k，其特征 $MSK^k = SP^k \times \text{Aap}$ 表示相应序列特征与尺度的乘积，其第 j 个元素 $MSK^{k,j} = SP^{k,j} \times \text{Aap}^j$，$j = 1,...,20$ 。使用 MSK^k 的标准差 TD^k 来表示第 k 个残基。因此，第 i 个目标蛋白质被向量化为 $TD = [TD^1,\cdots,\ TD^k,\cdots,\ TD^{lenSeq}]^T$，其中 lenSeq 是目标序列的长度。类似的向量表示形式可在我们的先前工作中找到 [31,33,34]。

2.2. 代表候选药物的特征向量

此外，为了对候选药物进行编码，使用了 PaDEL-Descriptor 软件。PaDEL-Descriptor 是一款用于计算分子描述符和指纹的软件，目前可计算不同的描述符（1D、2D 描述符和 3D 描述符）以及 10 种指纹[35]。分子描述符是将分子符号表示中编码的化学信息通过逻辑和数学过程转化为有用数字或某些标准化实验结果的最终产物[36]。在本研究中，使用了 1D 和 2D 描述符，同时从分子中去除盐分，假定最大的片段即为所需的分子。此外，在计算描述符之前，会自动检测分子中的芳香性信息并将其移除。所使用的 1D 和 2D 描述符列于下表 6 中。

因此，使用 1444 个描述符来编码一个药物分子。所以第 i 个候选药物可以表示为 $D^i = [D_1^i,\ D_2^i,\cdots,\ D_{1444}^i]^T$。这些 1D 和 2D 描述符以及指纹主要通过化学开发工具包 [35] 计算得出。这些描述符包括原子类型电拓扑状态描述符、麦高恩体积、分子线性自由能关系描述符、环计数以及由 Laggner [37] 确定的化学子结构计数。

对于第 i 对药物 - 目标组合 DT^i，其目标由 AAindex1 属性 Aap 编码，可以表示为一个（1444 + *lenSeq*）维的向量，其表达式为

$$V^{i,Aap} = [D^i,\ TD_A^i ap] = [D_1^i,\ D_2^i,\ldots,D_{1444}^i,\ TD_1^{i,Aap},\ TD^{i,Aap},\ldots,TD_{lenSeq}^{i,Aap}]^T, \tag{1}$$

其中 lenSeq 是目标序列的长度。

相应的目标值 T^i 为 1 或 0，表示药物 - 靶点对是否相互作用。实际上，我们的方法期望学习输入矩阵 V^{Aap} 与对应的目标数组 T 之间的关系，并努力使输出尽可能接近目标数组 T，其中 Aap 表示目标是基于不相关的 AAindex1 属性 Aap 进行编码的。

2.3. REPTree 上的随机投影

随机投影是一种将高维数据投影到低维子空间的数据降维技术[38-41]。给定原始数据向量 $X \in R^{N\times L1}$，线性随机投影是将原始向量乘以一个随机矩阵 $R \in R^{L1\times}$ L2。投影

$$X^R = XR \sum_i x_i r_i \tag{2}$$

得到一个降维后的向量 $X^R \in R^{N\times L2}$，其中 x_i 是原始数据的第 i 个样本，r_i 是随机矩阵的第 i 列，且 $L2 \ll L1$。矩阵 R 由随机值组成，且每列都已归一化为 1。在式（1）中，每个维度为 L_1 的原始数据样本在降维空间中都被替换为一个随机的、非正交的方向 L_2 [39]。因此，原始数据的维度从（1444 + *lenSeq*）降低到了一个相当小的值。

REPTree 是一种快速的树学习器，它采用基于信息增益/方差减少的减枝法[27]，并通过为数值属性排序来优化速度。本研究采用 REPTree 的默认 numFolds 参数（在 WEKA 软件中默认值为 3），该参数用于确定剪枝集的大小：数据被平均分成该数量的若干部分，最后一部分用作独立测试集，以估计每个节点的误差。

先前的研究结果表明，由一个分类器造成的泛化误差可以通过其他分类器来补偿，因此使用树集成能够显著提高预测的准确性[42]。对于药物-靶点相互作用预测问题，简单树的集成会为药物-靶点相互作用中最常见的类别投票。给定训练数据集 $Vtrk,Aap = \{(\text{XiR},_{\text{Aap}},\ Y_i)\}|_{iN=1}$ in $_k$

对于 AAindex1 属性 Aap，乘以随机投影 R^k 后，设训练实例的数量为 N，分类器中的特征数量为 L2。然后生成数据 $V^{k,Aap}$ 作为 REPTree 的输入，从而构建一个分类器 $CF_{k,Aap}(x)$，其中 x 是一个训练实例。

在所有采用随机投影的 REPTree 分类器生成完毕后，它们会为最热门的类别投票，因此集成模型的预测结果为：

$$Pred(X) = majority\ vote\ \{CF_{k,Aap}(x)\}_{Aap=1}^{34}, \tag{3}$$

其中 x 是一个查询实例。

结果表明，多数投票结合独立分类器通常能带来显著的改进[43,44]。这里，如果所有分类器都将一对药物 - 靶点识别为正类 1，则将其标记为相互作用，否则将其识别为不存在药物 - 靶点相互作用（见表 1）。

3. 材料

3.1. 数据集

我们在研究中使用了文献[12]中的药物靶点数据集。该数据集排除了缺乏实验信息的药物靶点对，最终包含 4797 对，其中针对酶的有 2719 对，针对离子通道的有 1372 对，针对 G 蛋白偶联受体（GPCR）的有 630 对，针对核受体的有 86 对。这些对的列表可在文献[12]中找到，详细信息可从 KEGG [45] 获取。在本研究中，所有这些数据集都被视为阳性数据集。

表 1
PaDEL-Descriptor 中使用的 1D 和 2D 描述符。

描述符类型	Number[a]	描述符类型	数字
酸性基团计数	1	巴里什矩阵	91
ALOGP	3	APol	1
芳香原子数	1	芳香键计数	1
原子数	14	自相关性	346
基本组计数	1	BCUT	6
键数	10	BPol	1
修正特征值	96	离心连接指数	1
碳类型	9	奇链	10
气团	8	气路簇	6
气道	32	宪法的	12
克里彭洛普和分子折射率	2	迂回矩阵	11
顶点邻接信息（数量）	1	原子类型电拓扑状态	489
FMF描述符	1	片段复杂度	1
氢键受体数	4	氢键供体数	2
杂交率	1	信息内容	42
卡帕形状指数	3	最大连锁店	1
最大的Pi系统	1	最长的脂肪链	1
曼恩霍尔德 LogP	1	麦高恩体积	1
分子距离边	19	分子线性自由能关系	6
路径计数	22	佩蒂让数	1
环数	68	可旋转键数	4
五规则	1	拓扑的	3
拓扑电荷	21	范德华体积	1
拓扑距离矩阵	11	拓扑极性表面积	1
步行计数	20	重量	2
加权路径	5	维纳数	2
XLogP	1	扎格列布指数	1
扩展拓扑化学原子	43		

每种类型中的描述符数量。

表 2
药物靶点数据集详情。

数据集	毒品	目标	互动 - 正面组合	负样本对
酶	419	643	2719	5438
离子通道	203	198	1372	2744
G 蛋白偶联受体	217	92	620	1240
核受体	53	25	86	172
总共	892	958	4797	9588

表 3
采用多数投票技术的 REPTree 分类器集成的预测性能，即集成系统预测药物 - 靶点对相互作用，当集成中的所有 REPTree 分类器都预测其相互作用时。

数据集	目标类型	Rec	Acc	Prec	F1
训练一个	酶	0.970	0.944	0.876	0.921
	离子通道	0.986	0.886	0.751	0.853
	G 蛋白偶联受体	0.994	0.892	0.758	0.860
	核受体	0.709	0.812	0.722	0.716
Test[b]	酶	0.972	0.900	0.782	0.867
	离子通道	0.993	0.89	0.755	0.858
	G 蛋白偶联受体	1.000	0.852	0.693	0.818
	核受体	0.837	0.911	0.889	0.862

a. 对训练数据集的预测 $\aleph_{tr}$。 b. 对测试数据集的预测 $\aleph_{ts}$。

表 4
在不同投影维度下，我们的方法在 Eq. (1)[a] 中 *L2* 的性能比较。

L2	目标类型	Rec	Acc	Prec	F1
3	酶	0.972	0.900	0.782	0.867
	离子通道	0.993	0.890	0.755	0.858
	G 蛋白偶联受体	1.000	0.852	0.693	0.818
	核受体	0.837	0.911	0.889	0.862
	平均值	0.951	0.888	0.780	0.851
5	酶	0.827	0.868	0.994	0.903
	离子通道	0.741	0.810	1.000	0.851
	G 蛋白偶联受体	0.384	0.527	0.971	0.550
	核受体	0.876	0.609	0.653	0.748
	平均值	0.707	0.704	0.905	0.762
10	酶	0.838	0.813	0.908	0.872
	离子通道	0.530	0.500	0.763	0.625
	G 蛋白偶联受体	0.849	0.461	0.541	0.661
	核受体	0.748	0.679	0.815	0.780
	平均值	0.741	0.613	0.757	0.735
20	酶	0.771	0.706	0.830	0.799
	离子通道	0.724	0.595	0.734	0.729
	G 蛋白偶联受体	0.781	0.700	0.815	0.797
	核受体	0.477	0.576	0.932	0.631
	平均值	0.688	0.644	0.828	0.739
50	酶	0.830	0.687	0.763	0.795
	离子通道	0.428	0.512	0.883	0.576
	G 蛋白偶联受体	0.465	0.566	0.930	0.620
	核受体	0.815	0.710	0.800	0.807
	平均值	0.635	0.619	0.844	0.700
100	酶	0.819	0.670	0.751	0.784
	在渠道中	0.737	0.724	0.882	0.803
	G 蛋白偶联受体	0.535	0.610	0.920	0.676
	核受体	0.739	0.694	0.842	0.787
	平均值	0.708	0.675	0.485	0.763

对测试数据集的预测 $\aleph_{ts}$

表 5
在相同的测试数据集上，我们的方法与另外两种方法在准确率方面的性能比较。

方法	类型	酶	离子通道	G 蛋白偶联受体	核受体
我们的方法	REPTree	0.900	0.890	0.852	0.911
参考文献 [12]	kNN	0.855	0.808	0.785	0.857
-药物		0.910[a]（此部分为数字和专有名词，无需翻译）	0.873 亿	0.855 倍光速	0.892 天
随机预测变量		0.489	0.489	0.488	0.488

关于 iEzy-Drug 预测器及其所报道的成功率，请参见参考文献 [48]。

b 关于 iCDI-Drug 预测器及其报告的成功率，参见参考文献 [49]。c 关于 iGPCR-Drug 预测器及其报告的成功率，参见参考文献 [50]。d 关于 iNR-Drug 预测器及其报告的成功率，参见参考文献 [51]。

算法 1. 通过随机投影预测药物 - 靶点相互作用：

Require: Training drug-target set $\aleph_{tr}$ and test set $\aleph_{ts}$ by applying 10-fold cross-validation technique to V of each type of drug-target interactions
Ensure: Prediction *Acc*
  for running times $1 \sim 100$ **do**
    Obtain a random projection R^{k};
    for AAindex1 property $Aap = 1 \sim 34$ **do**
  Run the REPTree classifier on $\aleph_{tr}^{k,Aap}$ by cross-validation;
  Obtain the prediction $Pred(X)^{k,Aap}$);
  end for

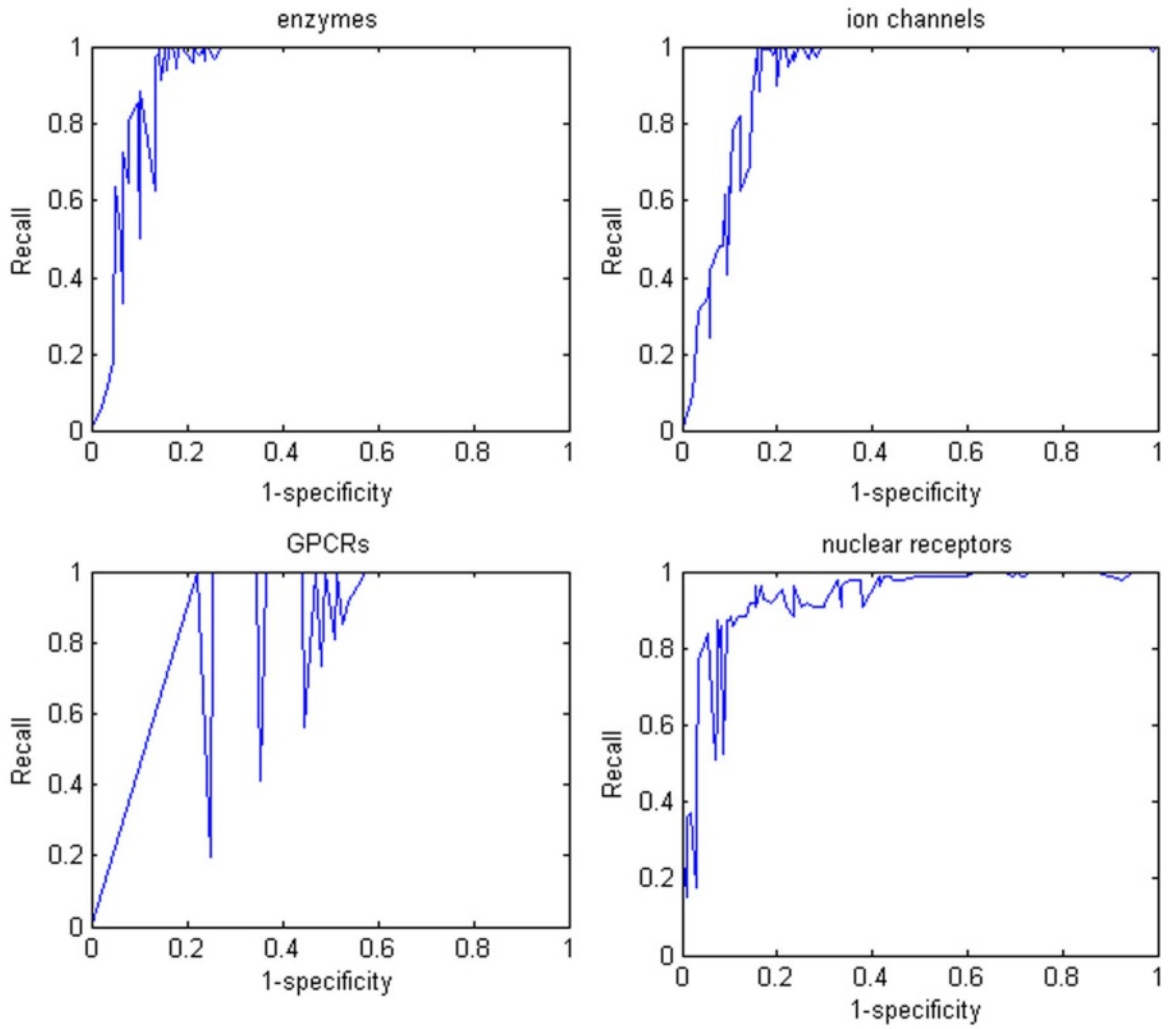

图 1. 我们的方法对酶类、离子通道、G 蛋白偶联受体和核受体各类别的受试者工作特征（ROC）性能。

Majority vote $Pred(X)^k$ to the predictions in terms of AAindex1 properties;
end for
Sort $Pred(X)$ and obtain the random projections with top accuracy;
Apply the top random projections to the test set $\aleph_{ts}$;
Calculate the performance on $\aleph_{ts}$.

相应的负样本数据集是通过参考文献[12]中的相同选择步骤获得的。选择步骤如下：（1）将上述正样本数据集中的配对拆分为单个药物和蛋白质；（2）将这些单个药物和蛋白质重新配对，确保它们在对应的正样本数据集中均未出现；（3）随机选取这样形成的负样本配对，直至其数量达到正样本配对数量的两倍[12]。药物-靶点相互作用对按照蛋白质靶点家族进行划分。总共 4797 个药物-靶点对被分为四个家族：酶、G 蛋白偶联受体（GPCRs）、离子通道和核受体。最终，四个数据集分别包含 8157、4116、1860 和 258 个酶、离子通道、GPCRs 和核受体的配对。表 2 列出了这四个数据集的详细信息。

3.2. 药物-靶点相互作用预测评估

在这项工作中，我们采用了四项评估指标来客观地展示我们模型的能力，即召回率（Rec）、准确率（Prec）、F 值（F1）和准确度（Acc）[33,46,47]。它们的定义如下：

$$Rec = \frac{TP}{TP+FN} Prec = \frac{TP}{TP+FP} Acc = \frac{TP+TN}{TP+FN+FP+FN} F1 = 2 \times \frac{Prec \times Sen}{Prec + Sen}, \quad (4)$$

其中，TP（真阳性）表示正确预测的药物 - 靶点对的数量；FP（假阳性）表示假阳性（错误地过度预测的非药物 - 靶点对）的数量；TN（真阴性）表示正确预测的非药物 - 靶点对的数量；FN（假阴性）表示假阴性，即错误地低估的药物 - 靶点对的数量。

4. 结果

4.1. 药物-靶点相互作用预测的性能

在本研究中，药物-靶点相互作用被分为四种类型：酶、离子通道、G 蛋白偶联受体（GPCRs）和核受体。对于每种类型，所提出的方法均予以应用（见算法 1），结果如下所示。每种类型药物-靶点相互作用的数据集通过 10 折交叉验证被分为训练数据集 $\aleph_{tr}$ 和测试数据集 $\aleph_{ts}$。也就是说，数据集被划分为 10 个大小大致相同的子集，每次将其中一个子集作为测试集，其余子集合并为训练集。依次选取测试子集，最终对所有实例进行测试。然后使用不同的随机投影将原始数据集投影到一个维度较低的空间，在本研究中为 5 维。

表 6
使用了 AAindex1 数据库中的 34 个属性。

增加	数据描述	类型
ARGP820101 ARGP820102	疏水性指数（阿戈斯等人，1982 年） 信号序列螺旋势能（阿戈斯等人，1982 年）	
ARGP820103	膜埋藏偏好参数（阿戈斯等人，1982 年）	疏水性
BULH740101	将自由能转移到表面（布尔 - 布里斯，1974 年）	
BULH740102	表观偏摩尔体积（Bull-Breese，1974）	
BIGC670101	残余体积（比格罗，1967 年）	
生物 V880101	信息价值用于可访问性；平均比例 35%（Biou 等人，1988 年）	
生物 V880102	信息价值用于可访问性；平均占比 23%（Biou 等人，1988 年）	残余体积
CHAM820101	极化率参数（查尔顿 - 查尔顿，1982 年）	
CHAM820102（此编号或	水溶液的自由能，千卡/摩尔（查顿 - 查顿，1982 年）	
CHOC750101	（乔西亚，1975 年）平均埋藏残基量	
CHOC760101	三肽中的残基可接触表面积（Chothia，1976）	
CHOC760102	折叠蛋白质中可接触的残基表面积（Chothia，1976）	
BHAR880101	平均柔韧性指数（Bhaskaran-Ponnuswamy，1988 年）	
BROC820101	三氟乙酸（TFA）中的保留系数（Browne 等人，1982 年）	灵活性
BROC820102	HFBA 中的保留系数（Browne 等人，1982 年）	
BEGF750101	内螺旋构象参数（贝金 - 德尔克克斯，1975 年）	
BEGF750102	β 结构的构象参数（贝金 - 迪尔克斯，1975 年）	
BEGF750103	β-转角的构象参数（贝金 - 德尔克克斯，1975 年）	
布拉740101	α-螺旋的归一化频率（Burgess 等人，1974 年）	
布拉740102	扩展结构的归一化频率（Burgess 等人，1974 年）	二级结构
查姆830101	螺旋构象的周-法斯曼参数（查尔顿-查尔顿，1983 年）	
CHAM830102（此编号或	由 Chou-Fasman β-折叠参数的最佳相关性所获得的残差定义的一个参数（Charton-Charton，1983）	
CHAM830103（此编号或	标记为 1+1 的侧链中的原子数（查尔顿 - 查尔顿，1983 年）	
查姆830104	标记为 2+1 的侧链中的原子数（查尔顿 - 查尔顿，1983 年）	
查姆830105	标记为 3+1 的侧链中的原子数（查尔顿 - 查尔顿，1983 年）	
查姆830106	最长链中的键数（查顿 - 查顿，1983 年）	
ANDN920101	α-C 原子的化学位移（安德森等人，1992 年）	
布纳790101	α-NH 化学位移（邦迪 - 维特里希，1979 年）	
布纳790102	α-CH 化学位移（邦迪 - 维特里希，1979 年）	
布纳790103	自旋 - 自旋耦合常数 3JHα-NH（邦迪 - 维特里希，1979 年）	立体参数
CHAM810101（此编号或	立体参数（查尔顿，1981 年）	
查姆830107	电荷转移能力的一个参数（查顿 - 查顿，1983 年）	
查姆830108	电荷转移给体能力的一个参数（查顿 - 查顿，1983 年）	

为了实现更好的随机投影，通过 10 折交叉验证将训练数据集 $\aleph_{tr}$ 分为训练子集 ℵtrsub 和测试子集 ℵtssub 。通过随机投影技术运行 REPTree 分类器，利用训练子集 ℵtrsub 对测试子集 ℵtssub 进行预测。仅保留性能最佳的随机投影。

给定最佳随机投影，对训练数据集 $\aleph_{tr}$ 和测试数据集 $\aleph_{ts}$ 运行 REPTree 分类器，从而得到最终预测结果。

具体而言，针对 34 个独立的 AAindex1 属性，对原始数据矩阵进行了 34 次随机投影 R^k 。通过随机投影集成的 34 个分类器对训练数据子集 ℵtrsub 和测试数据子集 ℵtssub 进行预测。如果预测准确率大于 0.75，则保留这 34 次随机投影 $\aleph_{tr}$ 。通过随机投影重复集成分类器 R^k，在 $\aleph_{tr}$ 和 $\aleph_{ts}$ 上通过随机投影 R^k 获得若干最佳预测（k = 1 ☒ K）。将 K 个预测结果合并，得出最终预测。表 3 展示了针对四个蛋白质靶点类别的集成性能比较。这里，原始数据的维度从（1444 + *lenSeq*）降低到 5。对于药物，它们被编码为固定长度 1444 的向量，而对于具有不同序列长度的蛋白质靶点，它们可以被编码为不同序列长度 lenSeq 的向量。随机投影的原始空间维度 maxLenSeq 采用最长的序列长度。序列长度较短的靶点序列在空间 $R^{maxLenSeq}$ 中被编码为子空间 RlenSeq，即 $R^{lenSeq} \in R^{maxLenSeq}$ 。从表 3 可以看出，在核受体类上测试的集成系统表现优于其他类别。在召回率为 0.837 时，其准确率为 0.911，精确率为 0.889。

4.2. 不同投影维度下的性能表现

我们采用随机投影技术来寻找优化的特征空间，并进一步将其应用于药物 - 靶点相互作用的预测。为此，研究了具有不同投影维度的随机投影。表 4 列出了根据方程（1）中不同 *L2* 的性能比较。从表 4 可以看出，投影维度 *L2* = 3 的实验表现优于其他实验，准确率达到 0.888。似乎投影维度较小的实验比投影维度较大的实验表现更好。

4.3. 与其他方法的比较

我们还将我们的方法与另外两种方法进行了比较：文献[12]中的工作以及在相同数据集上的随机预测器。表5展示了我们的方法与其他两种方法在准确率方面的性能比较。此处实现了随机预测器，并运行了100次。平均性能附在表的底部。对于酶类、离子通道、GPCR 和核受体这四类，我们的方法分别达到了 0.900、0.89、0.852 和 0.911 的准确率。与文献[12]中的工作相比，我们的方法在这四类上的准确率提高了 4.5% 至 8.2%。此外，我们的方法与四个网络服务器：iEzy-Drug、iCDI-Drug、iGPCR-Drug 和 iNR-Drug 的表现相当。而且结果表明，我们的方法的准确率是随机预测器的两倍。

图 1 展示了采用多数投票法的集成分类器的性能。尽管在 GPCRs 类中识别药物 - 靶点对要困难得多，但我们的方法仍能得出良好的预测结果。

4.4. 34 处房产描述

由于从 AAindex1 数据库中提取了 34 个独立的氨基酸属性并将其应用于药物靶点预测，因此这 34 个属性的详细信息列于表 6 中。表中展示了每个氨基酸属性的数据描述。其中一些属性用于疏水性指数，一些用于残基体积，一些用于柔性指数，一些用于二级结构，还有一些用于原子间相互作用。这些氨基酸属性对于蛋白质序列编码十分重要，因为它们通过不同的环境特征来表示蛋白质序列。编码方案旨在应用

多种统计特征用于恢复氨基酸残基之间的真实相互作用。

5. 结论

本文提出了一种通过随机投影的 REPTree 分类器集成来识别药物 - 靶点相互作用的方法。对于每个独立的 AAindex1 属性，将通过不同随机投影变换的药物 - 靶点相互作用的原始编码器输入到一个 REPTree 分类器中。针对 AAindex1 属性共有 34 个 REPTree 分类器。这些 REPTree 分类器的集成能够对药物 - 靶点相互作用做出良好的预测。因此，我们的方法简单，仅应用了统计氨基酸属性。此外，这里采用了随机投影的降维技术来减少原始编码器空间。更重要的是，随机投影技术能够处理具有不同氨基酸数量的蛋白质链，并获得统一的编码器空间。实际上，随机投影技术提供了一种有用的机制，它降低了高维原始数据的维度，使数据更加多样化，从而该方法在药物 - 靶点相互作用的预测中表现良好。结果表明，我们的方法在药物 - 靶点相互作用的预测中优于其他最先进的方法。

致谢

本研究得到了国家自然科学基金（项目编号：61672035、61300058、61472282 和 61271098）的支持。

参考文献

[1] J. Knowles, G. Gromo, A guide to drug discovery, target selection in drug discovery, Nat. Rev. Drug Discov. 2 (1) (2003) 63–69. http://dx.doi.org/10.1038/nrd986.

[2] Johnson, Wolfgang, Predicting human safety screening and computational ap-proaches, Drug Discov. Today 5 (10) (2000) 445–454.

[3] S. Sirois, G. Hatzakis, D. Wei, Q. Du, K.-C. Chou, Assessment of chemical libraries for their druggability, Comput. Biol. Chem. 29 (1) (2005) 55–67. http://dx.doi.org/10.1016/j.compbiolchem.2004.11.003.

[4] A.J. Wood, W.E. Evans, H.L. McLeod, Pharmacogenomics drug disposition, drug targets, and side effects, New Engl. J. Med. 348 (6) (2003) 538–549.

[5] J.-F. Wang, D.-Q. Wei, C. Chen, Y. Li, K.-C. Chou, Molecular modeling of two cyp2c19 SNPs and its implications for personalized drug design, Protein Pept. Lett. 15 (1) (2008) 27–32.

[6] J.-F. Wang, D.-Q. Wei, K.-C. Chou, Pharmacogenomics and personalized use of drugs, Curr. Top. Med. Chem. 8 (18) (2008) 1573–1579.

[7] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, Y. Yamanishi, Relating drug-protein interaction network with drug side effects, Bioinformatics 28 (18) (2012) i522–i528.

[8] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (3) (1996) 470–489. http://dx.doi.org/10.1006/jmbi.1996.0477.

[9] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E. S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, Nat. Biotechnol. 25 (1) (2007) 71–75. http://dx.doi.org/10.1038/nbt1273.

[10] S. Zhu, Y. Okuno, G. Tsujimoto, H. Mamitsuka, A probabilistic model for mining implicit chemical compound-generelations from literature, Bioinformatics 21 (Suppl. 2) (2005) ii245–ii251.

[11] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (13) (2008) i232–i240. http://dx.doi.org/10.1093/bioinformatics/btn162.

[12] Z. He, J. Zhang, X.-H. Shi, L.-L. Hu, X. Kong, Y.-D. Cai, K.-C. Chou, Predicting drug-target interaction networks based on functional groups and biological features, PLoS One 5 (3) (2010) e9603. http://dx.doi.org/10.1371/journal.pone.0009603.

[13] Y.-C. Wang, C.-H. Zhang, N.-Y. Deng, Y. Wang, Kernel-based data fusion improves the drug-protein interaction prediction, Comput. Biol. Chem. 35 (6) (2011) 353–362.

[14] N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, H. Kobayashi, M. Imoto, Y. Sakakibara, Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening, PLoS Comput. Biol. 5 (6) (2009) e1000397.

[15] K.-C. Chou, A vectorized sequence-coupling model for predicting hiv protease cleavage sites in proteins, J. Biol. Chem. 268 (23) (1993) 16938–16948.

[16] X. Xiao, P. Wang, K.-C. Chou, Gpcr-ca: a cellular automaton image approach for predicting g-protein-coupled receptor functional classes, J. Comput. Chem. 30 (9) (2009) 1414–1423.

[17] K.-C. Chou, H.-B. Shen, Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2) (2008) 153–162.

[18] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, Predict drug-protein interaction in cellular networking, Curr. Top. Med. Chem. 13 (14) (2013) 1707–1712.

[19] K.-C. Chou, D.W. Elrod, Prediction of membrane protein types and subcellular locations, Proteins: Struct. Funct. Bioinform. 34 (1) (1999) 137–153.

[20] K.-C. Chou, H.-B. Shen, et al., Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (02) (2009) 63.

[21] L. Zhu, Z.-H. You, D.-S. Huang, B. Wang, t-lse: a novel robust geometric approach for modeling protein-protein interaction networks, PLoS One 8 (4) (2013) e58368. http://dx.doi.org/10.1371/journal.pone.0058368.

[22] D.-S. Huang, H.-J. Yu, Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2) (2013) 457–467. http://dx.doi.org/10.1109/TCBB.2013.10.

[23] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, H. Zhang, Prediction of protein-protein interactions based on protein-protein correlation using least squares regression, Curr. Protein Pept. Sci. 15 (6) (2014) 553–560.

[24] B. Wang, D.-S. Huang, C. Jiang, A new strategy for protein interface identification using manifold learning method, IEEE Trans. Nanobiosci. 13 (2) (2014) 118–123. http://dx.doi.org/10.1109/TNB.2014.2316997.

[25] L. Zhu, S.-P. Deng, D.-S. Huang, A two-stage geometric method for pruning unreliable links in protein-protein networks, IEEE Trans. Nanobiosci. 14 (5) (2015) 528–534. http://dx.doi.org/10.1109/TNB.2015.2420754.

[26] S.-P. Deng, L. Zhu, D.-S. Huang, Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks, BMC Genom. 16 (Suppl. 3) (2015) S4. http://dx.doi.org/10.1186/1471-2164-16-S3-S4.

[27] F. Esposito, D. Malerba, G. Semeraro, V. Tamma, The Effects of Pruning Methods on the Predictive Accuracy of Induced Decision Trees, 1999.

[28] X.Z. Fern, C.E. Brodley, Random projection for high dimensional data clustering: a cluster ensemble approach, in: ICML, vol. 3, 2003, pp. 186–193.

[29] A. Schclar, L. Rokach, Random projection ensemble classifiers, in: Enterprise Information Systems, Springer, 2009, pp. 309–316.

[30] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, Aaindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205. http://dx.doi.org/10.1093/nar/gkm998(Database issue).

[31] P. Chen, J. Li, L. Wong, H. Kuwahara, J.Z. Huang, X. Gao, Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences, Proteins 81 (8) (2013) 1351–1362. http://dx.doi.org/10.1002/prot.24278.

[32] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

[33] P. Chen, J. Li, Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information, BMC Bioinform. 11 (2010) 402. http://dx.doi.org/10.1186/1471-2105-11-402.

[34] P. Chen, L. Wong, J. Li, Detection of outlier residues for improving interface prediction in protein heterocomplexes, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (4) (2012) 1155–1165. http://dx.doi.org/10.1109/TCBB.2012.58.

[35] C.W. Yap, Padel-descriptor: an open source software to calculate molecular descriptors and fi ngerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474. http://dx.doi.org/10.1002/jcc.21707.

[36] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, vol. 11, John Wiley & Sons, 2008.

[37] J. Klekota, F.P. Roth, Chemical substructures that enrich for biological activity, Bioinformatics 24 (21) (2008) 2518–2525. http://dx.doi.org/10.1093/bioinfor-matics/btn479.

[38] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent Semantic Indexing: a Probabilistic Analysis, 1998.

[39] S. Kaski, Dimensionality reduction by random mapping: fast similarity computa-tion for clustering, in: Proceedings of the Neural Networks, IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, vol. 1, 1998, pp. 413–418 http://dx.doi.org/10.1109/IJCNN.1998.682302 〈http://ieeexplore.ieee.org/stamp/stamp.jsp?Arnumber=682302〉.

[40] Z. Wang, W. Jie, S. Chen, D. Gao, Random projection ensemble learning with multiple empirical kernels, Knowl. Based Syst. 37 (2013) 388–393.

[41] A. Ahmad, G. Brown, Random projection random discretization ensembles– ensembles of linear multivariate decision trees, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1225–1239. http://dx.doi.org/10.1109/TKDE.2013.134 〈http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6574846〉.

[42] P. Chen, J.Z. Huang, X. Gao, LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone, BMC Bioinform. 15 (Suppl. 15) (2014) S4. http://dx.doi.org/10.1186/1471-2105-15-S15-S4.

[43] P. Chen, J. Li, Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers, BMC Struct. Biol. 10 (Suppl. 1) (2010) S2. http://dx.doi.org/10.1186/1472-6807-10-S1-S2.

[44] L.I. Kuncheva, C.J. Whitaker, R.P.W. Duin, Limits on the Majority Vote Accuracy in Classifi er Fusion, 2003.

[45] M. Kanehisa, The Kegg Database, Novartis Found. Symp., vol. 247, 2002 91–101; discussion 101–3, 119–28, 244–52.

[46] P. Chen, C. Liu, L. Burge, J. Li, M. Mohammad, W. Southerland, C. Gloster, B. Wang, DomSVR: domain boundary prediction with support vector regression

from sequence information alone, Amino Acids 39 (3) (2010) 713–726. http://dx.doi.org/10.1007/s00726-010-0506-6.

[47] B. Wang, P. Chen, D.-S. Huang, J.-j. Li, T.-M. Lok, M.R. Lyu, Predicting protein interaction sites from residue spatial sequence profile and evolution rate, FEBS Lett. 580 (2) (2006) 380–384. http://dx.doi.org/10.1016/j.febslet.2005.11.081.

[48] J.-L. Min, X. Xiao, K.-C. Chou, Iezy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking, Biomed. Res. Int. 2013 (2013) 701317. http://dx.doi.org/10.1155/2013/701317.

[49] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, iCDI-PseFpt: identify the channel-drug interaction in cellular networking with pseaac and molecular fingerprints, J. Theor. Biol. 337 (2013) 71–79. http://dx.doi.org/10.1016/j.jtbi.2013.08.013.

[50] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, igpcr-drug: a web server for predicting interaction between gpcrs and drugs in cellular networking, PLoS One 8 (8) (2013) e72234. http://dx.doi.org/10.1371/journal.pone.0072234.

[51] Y.-N. Fan, X. Xiao, J.-L. Min, K.-C. Chou, Inr-drug predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. 15 (3) (2014) 4915–4937. http://dx.doi.org/10.3390/ijms15034915.

彭晨专攻机器学习和数据挖掘，其研究成果应用于生物信息学、药物发现、计算机视觉等领域。他在国际会议和期刊上发表了约 40 篇高质量的学术论文。他是中国合肥安徽大学生命科学学院的教授。他拥有电子工程学院（中国人民解放军）的学士学位、昆明理工大学的硕士学位以及中国科学技术大学的博士学位。在加入安徽大学之前，他曾任职于香港城市大学（2006 年，担任高级研究助理）、美国霍华德大学（2008 - 2009 年，担任博士后研究员）、南洋理工大学。新加坡科技大学（2009 - 2010 年，任研究员），沙特阿拉伯阿卜杜拉国王科技大学（2012 - 2014 年，任博士后研究员）。2011 年至 2013 年，他在中国科学院合肥智能机械研究所担任副教授。

朱慕春是安徽大学生命科学学院的一名硕士研究生。他的研究方向是生物信息学和软件应用。

张军 1971 年出生于中国安徽省。2004 年，他在中国科学院智能机械研究所获得模式识别与智能系统硕士学位。2007 年，他在中国科学技术大学（位于中国合肥）获得博士学位。目前，张博士是中国安徽大学电气工程与自动化学院的副教授。他的研究兴趣集中在深度学习、集成学习和化学信息学方面。

王兵于 1998 年和 2004 年分别获得合肥工业大学的学士和硕士学位，2006 年获得中国科学技术大学博士学位。目前，王博士在中国上海同济大学电子与信息工程学院担任研究教授。他的研究兴趣主要集中在机器学习、计算生物学和化学信息学方面。