基于网络和基因差异表达信息的癌症致病基因预测

鱼亮*, 任姝洁

西安电子科技大学计算机科学与技术学院, 西安 710071

* 联系人, E-mail: lyu@xidian.edu.cn

收稿日期: 2022-06-23; 接受日期: 2022-07-06; 网络版发表日期: 2022-09-29

国家重点研发计划(批准号: 2018YFC0910403)和国家自然科学基金项目(批准号: 62072353)资助

摘要 疾病致病基因的发现是基因组研究的重大挑战之一. 近年, 随着生物学数据的积累, 许多研究人员利用计算方法进行致病基因预测. 但其中大多都基于基因相互作用网络或其他相似性网络数据等, 很少考虑特定基因的局部网络连接与它们的差异表达信息之间的潜在联系. 本文基于基因相互作用局部网络结构和基因的差异表达信息探索癌症致病基因及其邻居基因的生物特性, 并依据新发现的特性采用机器学习方法进行癌症致病基因预测. 首先, 从TCGA(The Cancer Genome Atlas)数据库和OMIM(Online Mendelian Inheritance in Man)数据库中获取21种癌症相关的基因表达数据及其致病基因数据, 并依次将人类蛋白质相互作用网络和各癌症对应的组织特异性相互作用网络作为背景网络, 分析不同生物学网络的邻域信息和患病前后基因表达的变化信息之间的潜在生物特性. 接着基于发现的生物特性定义基因节点特征的向量表示方法, 并采用支持向量机进行致病基因预测. 实验结果通过ICGC(International Cancer Genome Consortium), COSMIC(Catalogue Of Somatic Mutations In Cancer), NCG(Network of Cancer Genes), OncoKB(Oncology Knowledge Base)等标准数据库和相关文献, 以及疾病注释和通路富集进行验证. 结果表明, 根据发现的致病基因的特性进行基因特征定义, 能够将癌症致病基因与其他基因进行区分, 并为癌症致病基因预测提供有力假设, 为相关生物实验提供可靠致病基因候选集, 进而推动对癌症这一复杂疾病致病机理的研究.

关键词 癌症致病基因, 相互作用网络, 差异表达信息, 特征的向量表示, 支持向量机

挖掘疾病的致病基因, 不仅对了解疾病的发病机理和作用机制有着重要的意义^[1], 同时也对疾病的早期预防^[2]、临床诊断和治疗^[3]有着很大的帮助^[4]. 癌症作为人类健康的最大杀手之一, 目前临床上治疗的方法依旧为手术切除、化疗、放疗、干细胞移植、免疫

疗法等^[5]. 事实上, 突破癌症的关键点在于突破早期诊断的门槛^[6]. 对癌症致病基因的研究可以帮助人们发现不同癌症间更多的共性和特定癌症所具有的特性, 并以此为基础来推动抗癌药物的研究, 进而推动癌症早期诊断的实现.

引用格式: 鱼亮, 任姝洁. 基于网络和基因差异表达信息的癌症致病基因预测. 中国科学: 生命科学, 2023, 53: 94–108

Yu L, Ren S J. Prediction of cancerous pathogenic genes based on network and gene differential expression information (in Chinese). Sci Sin Vitae, 2023, 53: 94–108, doi: [10.1360/SSV-2021-0324](https://doi.org/10.1360/SSV-2021-0324)

随着生物信息学的出现和大量基因组数据的积累, 研究人员开始关注基因对疾病贡献的排序算法^[7], 并利用计算机技术从海量数据中挖掘候选致病基因. 按照实现手段, 常见的致病基因预测方法可被分为两大类: 统计分析方法和机器学习方法. 统计分析方法主要基于基因表达信息^[8]、基因功能注释、基因的序列信息^[9], 以及疾病相关基因的统计特征和已知的生物医学文献^[10]来进行致病基因预测, 从大量体细胞突变中寻找致病性^[11]. 例如定义评价函数分别对病例组和对对照组分析差异最大的单个核苷酸多态位点以此来获得致病位点等^[12]. 或者针对某种疾病, 对高通量测序所得的全基因组数据进行匹配对照, 在人群中进行大型遗传学分析来鉴定致病性基因及其遗传贡献^[13]. 此外, 也有相关方法致力于对致病基因模块的挖掘, Li 等人^[14]提出了一种高维大数据基因统计的致病性模块信息识别方法. 相比于统计学方法, 传统机器学习和深度学习方法由于其很高的预测性能受到了更加广泛的关注. 例如, Collier 等人^[15]提出的用于预测癌症驱动基因的单任务和多任务机器学习算法 LOTUS, 以及 Luo 等人^[16]提出的使用深度卷积神经网络基于体细胞突变预测癌症驱动基因的 deepDriver 方法. 此类方法还会基于某些生物学网络, 学习网络的局部信息和全局信息, 在完成异质信息融合后, 使用传统机器学习模型进行训练和预测^[17~19]. 但这类方法的可解释性往往不高, 难以给出特征的具体含义, 进而难以说明研究结果的生物学意义.

受上述两类方法的启发, 本文采用统计分析的方式探索生物学网络的邻域信息和患病前后基因表达的变化信息之间的潜在联系或某些较强隐含的规律. 并将此规律作为核心思想, 设计基因节点特征的向量表示方法, 采用机器学习中常见的分类器根据所设计的特征进行癌症致病基因的预测, 进一步验证猜想规律的合理性. 在先前研究中, 有科研人员将生物学网络的邻域信息和患病前后基因表达的变化信息作为指导来预测禾谷镰刀菌(*Fusarium graminearum*)的致病基因^[20]. 该方法利用少量已知的致病基因作为种子基因, 可以从禾本科镰刀菌的蛋白质-蛋白质相互作用网络中识别出一个由潜在致病基因组成的子网, 进一步要求子网中的基因在病原真菌入侵前后具有显著的差异表达. 本文将该思想从农作物迁移至人类, 为癌症致病基因预测寻找更多方案. 如图1所示, 本文根据已有数

据进行差异表达分析(differential expression analysis, DEA), 总结癌症致病基因的差异表达特性与其在基因-基因生物学网络中的邻接基因的差异表达特性, 依据观测所得的特性设计基因节点特征的向量表示方法, 最后基于该特征选取分类器完成癌症致病基因预测的任务.

1 材料与方法

1.1 数据收集

21种有明确致病基因的癌症数据. 癌症表达数据来源于The Cancer Genome Atlas(TCGA)数据库^[21], 癌症的致病基因来源于Online Mendelian Inheritance in Man(OMIM)数据库^[22], 所选癌症满足以下要求: (i) 表达数据包含患病后的实验组和患病前的对照组, 对照组的个数大于1, 以防止单个样本带来的随机误差; (ii) 在OMIM数据库中可以找到明确的与表型相关的致病基因. 选择的癌症分别为膀胱尿路上皮癌(Bladder Cancer, BLCA)、乳腺浸润癌(Breast Cancer, BRCA)、宫颈鳞癌和腺癌(Cervical Cancer, CESC)、结肠癌(Colon Cancer, COAD)、食管癌(Esophageal Cancer, ESCA)、多形成性胶质细胞瘤(Glioblastoma, GBM)、头颈鳞状细胞癌(Head and Neck Cancer, HNSC)、肾嫌色细胞瘤(Kidney Chromophobe, KICH)、肾透明细胞癌(Kidney Clear Cell Carcinoma, KIRC)、肾乳头状细胞癌(Kidney Papillary Cell Carcinoma, KIRP)、肝细胞肝癌(Liver Cancer, LIHC)、肺腺癌(Lung Adenocarcinoma, LUAD)、肺鳞癌(Lung Squamous Cell Carcinoma, LUSC)、胰腺癌(Pancreatic Cancer, PAAD)、嗜铬细胞瘤和副神经节瘤(Pheochromocytoma & Paraganglioma, PCPG)、前列腺癌(Prostate Cancer, PRAD)、直肠腺癌(Rectal Cancer, READ)、肉瘤(Sarcoma, SARC)、胃癌(Stomach Cancer, STAD)、甲状腺癌(Thyroid Cancer, THCA)、子宫内膜癌(Endometrioid Cancer, UCEC).

10种无明确致病基因的癌症数据. 其中1种癌症的表达数据来自TCGA数据库, 9种来自GEO数据库^[23], 且满足以下要求: (i) 表达数据包含患病后的实验组和患病前的对照组, 对照组的个数大于1, 以防止单个样本带来的随机误差; (ii) 可进行测序芯片标识与基因ID转换; (iii) 不同于训练集的21种癌症. 所

图 1 方法流程图
Figure 1 Flow chart of our pathogenic genes prediction method

选癌症分别为肾上腺皮质癌(Adrenocortical Cancer, ACC)、胆管癌(Bile Duct Cancer, CHOL)、恶性胸膜间皮瘤(Malignant Pleural Mesothelioma, MPM)、弥漫性恶性胸膜间皮瘤(Diffuse Malignant Pleural Mesothelioma, MPMD)、鼻咽癌(Nasopharyngeal Carcinoma, NACA)、成神经细胞瘤(Neuroblastoma, NB)、口腔癌(Oral Cancer, ORCA)、骨肉瘤(Osteosarcoma, OSTE)、原始神经外胚层肿瘤(Primitive Neuroectodermal Tumour, PNET)、舌癌(Tongue Carcinoma, TNCA)。

人类蛋白质相互作用网络. 采用整合了7种数据类型的网络, 包括来自于IntAct^[24]和MINT^[25]数据库的二元分子相互作用(binary), BioGRID^[26]及HPRD^[27]数据库提供的文献验证相互作用(literature), NCBI数据库的信号传导相互作用(signaling), CORUM^[28]数据库提供的蛋白质复合体(complexes), KEGG^[29]和BIGG^[30]数据库中获得的代谢中酶偶联相互作用(metabolic), TRANSFAC^[31]数据库提供的调控相互作用(regulatory), PhosphositePlus^[32]数据库中的激酶-基质对(kinase). 该网络包含13460个节点, 141296条边, 去掉节点的自环后得到的蛋白质相互作用网络包含13397个节点, 138428条边。

人类组织特异性相互作用网络. 基因的精确作用通常取决于其所在的组织环境, 因此, 本文从GIANT数据库^[33]收录的145个组织特异性相互作用网络中, 为21种有明确致病基因的癌症挑选相对应的网络. 其中19种癌症找到了明确匹配的组织特异性相互作用网络, 网络中节点数均大于25000。

1.2 预处理方法

为了消除不同癌症的差异表达分析结果之间的一致性, 选取最大最小归一化(max-min法)和标准化分数归一化(z-score法)对每个癌症的差异表达分析结果中的 $|\log FC|$ 值进行归一化. max-min的定义为

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)}, \quad (1)$$

其中, X^* 表示 X 值归一化后的结果, $\min(X)$ 和 $\max(X)$ 分别表示取 X 的最小值和最大值. z-score的定义为

$$X^* = \frac{X - \mu}{\sigma}, \quad (2)$$

其中, X^* 表示 X 值归一化后的结果, μ 表示所有 X 值的均值, σ 表示所有 X 值的方差。

1.3 分类模型

支持向量机(support vector machines, SVM)模型^[34]结构精简, 可以解决非线性、高维度问题, 选择使用该分类模型能够突出在预测任务中特征占主导地位, 而非模型本身. 基于SVM分类模型可以说明, 本文提出的两种癌症致病基因特性在后续研究具有指导作用。

1.4 评价指标

模型分类性能的评价指标为精度值(Precision)和AUC值(AUC-score), 相关定义如下:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (3)$$

$$\text{TPR} = \frac{TP}{TP + FN}, \quad (4)$$

$$\text{FPR} = \frac{FP}{FP + TN}, \quad (5)$$

其中, TP表示正确预测的致病基因数量, TN表示正确预测的非致病基因数量, FP表示错误预测为致病基因的非致病基因数量, FN表示错误预测为非致病基因的致病基因数量. 将FPR(false positive rate)和TPR(true positive rate)分别作为横、纵坐标绘制ROC曲线, 可得到ROC曲线下方的面积, 即AUC值, 来判断二分类预测模型的性能. 上述评价指标已在许多相关研究中被使用^[15-19]。

2 结果

2.1 癌症致病基因的差异表达特性

分析21种有明确致病基因的癌症所对应的致病基因和患癌前后差异表达基因的重叠情况. 筛选差异表达基因的指标: 矫正后的 P 值(false discovery rate, FDR)小于0.1, 倍数变化 $|\log FC|$ 大于1. 统计可得, 21种癌症中有19种对应的致病基因中非差异表达基因所占的比例大于0.6, 其中BLCA, HNSC, KIRC, PAAD, PRAD, THCA在OMIM数据库中记录的致病基因全部非差异表达. 但对于某些癌症, 如CESC, ESCA, GBM, 以上现象并不明显. 特别是CESC, 它在OMIM数据库

中被记录的致病基因过少, 是其结果呈现反向趋势的原因之一。由此可知, 癌症的致病基因在患病前后的表达量往往不具有显著差异或者差异不明显。因此, 推测癌症致病基因第一个特性: 致病基因患病前后表达倾向于无差异或者差异较小。

2.2 癌症致病基因的邻居基因的差异表达特性

为了探索癌症致病基因在不同生物网络下的邻域信息和患病前后基因表达的变化信息之间的联系, 本文选取了两类网络进行实验, 得到以下结果。

(1) 基于人类蛋白质相互作用网络. 将21种有明确致病基因的癌症对应的致病基因依次投影至蛋白质相互作用网络上, 分析致病基因一阶邻居的差异表达情况, 结果如表1所示。参数设置为: FDR小于0.1且倍数

变化 $|\log FC|$ 大于1。由结果可知, 以该蛋白质相互作用网络作为背景网络时, 有两种癌症GBM和LUSC, 其所有致病基因的一阶邻居中差异表达基因所占比例大于1/3(GBM: 0.3518; LUSC: 0.3457), 该比例远大于致病基因一阶邻居占网络中其余基因(不包括致病基因)的比例(GBM: 0.0481; LUSC: 0.0514)。即有1/3以上的差异表达基因聚集在以癌症致病基因为中心的1/20 (GBM: 0.0481; LUSC: 0.0514)的邻域中, 而其余近2/3的差异表达基因分散在网络其余19/20区域内。对于大多数癌症来说, 两个比例(表1的第6和7列)的差距较为明显, 其中11种癌症(表1加粗部分)的比例差距达到5倍以上, 该差距越大, 说明在此背景网络中癌症的差异表达基因在癌症致病基因周围的聚集程度越高。但对于某些癌症, 如SARC和PAAD, 以上现象并不明显。

表 1 癌症致病基因一阶邻居在人类蛋白质-蛋白质相互作用(Protein-Protein Interaction, PPI)网络中的差异表达情况

Table 1 Differential expression of first-order neighbors of cancerous pathogenic genes in the human PPI network

No.	癌症名称	致病基因个数	致病基因一阶邻居个数	一阶邻居中差异表达基因个数	一阶邻居中差异表达基因占所有一阶邻居的比例	一阶邻居中差异表达基因占整个网络中除致病基因以外所有基因的比例
1	GBM	6	651	229	0.3518	0.0481
2	LUSC	11	700	242	0.3457	0.0514
3	UCEC	4	149	43	0.2886	0.0108
4	BLCA	4	302	74	0.2450	0.0222
5	CESC	1	42	10	0.2381	0.0031
6	COAD	37	2048	460	0.2246	0.1501
7	READ	31	1680	375	0.2232	0.1231
8	PCPG	12	458	96	0.2096	0.0333
9	LUAD	14	798	167	0.2093	0.0585
10	KICH	7	321	58	0.1807	0.0234
11	BRCA	25	1537	275	0.1789	0.1129
12	KIRP	9	410	64	0.1561	0.0299
13	ESCA	6	98	15	0.1531	0.0069
14	STAD	12	607	91	0.1499	0.0444
15	KIRC	6	319	46	0.1442	0.0234
16	LIHC	9	1008	144	0.1429	0.0746
17	HNSC	2	69	8	0.1159	0.0515
18	THCA	9	218	24	0.1101	0.0007
19	PRAD	14	521	48	0.0921	0.0378
20	SARC	9	808	42	0.0520	0.0596
21	PAAD	9	865	10	0.0116	0.0639

a) 粗体表示比例差距达到5倍以上

本节实验中, 21种癌症筛选出的差异表达基因个数的平均值约为4500, SARC和PAAD在给定阈值下筛选出的差异表达基因个数(SARC: 1005; PAAD: 531)远低于平均值, 是其结果呈现反向趋势的原因之一.

(2) 基于人类组织特异性相互作用网络. 对于每种有明确致病基因的癌症, 依次取其对应的组织特异性相互作用网络作为背景网络, 进行与前文相似的分析, 分析结果如表2所示. GIANT数据库中组织特异性相互作用网络的相互作用关系来源很广, 单个网络的连通性通常较大, 不同网络之间稠密程度的差距较大. 仅关注表2中的数据, 网络稠密程度的差距最大可达3~4倍. 进而高的连通性又导致了致病基因一阶邻居规模庞大. 所以, 此处不再关注致病基因一阶邻居中差异表达基因所占的比例, 而是关注差异表达基因中一阶邻居的比例. 分析可得, 在19种癌症中, 有10种癌症(表2加粗部分)的差异表达基因有50%以上都是对应

癌症的致病基因在特定背景网络下的一阶邻居. 其中有5种癌症的差异表达基因有高达80%以上都是对应癌症的致病基因在特定背景网络下的一阶邻居.

综上所述, 无论是在人类蛋白质相互作用网络中, 还是在特定癌症对应的组织特异性相互作用网络中, 患病前后的差异表达基因倾向于聚集在癌症致病基因的周围. 因此, 推测癌症致病基因第二个特性: 癌症致病基因的一阶邻居在患病前后的表达值倾向于具有显著差异.

2.3 癌症致病基因预测

通过上述分析, 本文推测癌症致病基因具有两个特性: (i) 癌症致病基因患病前后表达值倾向于无差异或者差异较小; (ii) 癌症致病基因的一阶邻居在患病前后的表达值倾向于具有显著差异. 本节将基于以上两个特性构造表示基因节点特征的向量, 并将人类

表 2 癌症致病基因一阶邻居在人类组织特异性网络中的差异表达情况

Table 2 Differential expression of first-order neighbors of cancerous pathogenic genes in the human tissue-specific network

No.	癌症名称	组织特异性网络名称	网络规模(边数)	致病基因一阶邻居个数	一阶邻居中差异表达基因个数	癌症的差异表达基因个数	癌症的差异表达基因中是一阶邻居的比例
1	BLCA	Urinary bladder	215812287	25102	4416	4855	0.9096
2	GBM	Glia	203276776	25140	6925	7718	0.8973
3	PCPG	Neuron	126769446	24235	4447	5130	0.8669
4	LIHC	Hepatocyte	129962987	22493	3233	3880	0.8332
5	BRCA	Mammary epithelium	114461328	22886	3812	4718	0.8080
6	STAD	Stomach	91921490	20945	2660	3876	0.6863
7	CESC	Uterine cervix	89798723	15837	3270	5316	0.6151
8	KIRP	Renal tubule	103088769	18596	3017	5144	0.5865
9	KICH	Renal tubule	103088769	17468	3594	6587	0.5456
10	ESCA	Esophagus	120460773	17047	2097	3870	0.5419
11	THCA	Thyroid gland	83728113	16126	1488	3013	0.4939
12	KIRC	Renal tubule	103088769	16528	2790	5705	0.4890
13	COAD	Colon	65271825	15784	2096	5149	0.4071
14	PRAD	Prostate gland	66884562	13686	820	2549	0.3217
15	LUSC	Lung	59798192	11432	2061	7189	0.2867
16	LUAD	Lung	59798192	11812	1312	4910	0.2672
17	UCEC	Uterine endometrium	77194064	8268	1149	5627	0.2042
18	HNSC	Lymph node	80331157	8234	796	3993	0.1993
19	PAAD	Pancreas	64458733	12399	102	531	0.1921

a) 粗体表示癌症的差异表达基因中是一阶邻居的比例在50%以上

蛋白质相互作用网络作为背景网络, 从癌症的差异表达数据中提取样本, 最终展示通过SVM分类模型对癌症致病基因进行预测的结果.

(1) 基因特征的向量表示. 每个基因节点特征的向量表示信息由该基因节点(将其称为中心基因节点)以及它的一阶邻居信息构成, 共包含5个特征, 如图2所示. 特征1: 中心基因的 $|\log FC|$ 值, 用来反映该基因的差异表达程度; 特征2: 在人类蛋白质相互作用网络上, 以中心基因向外扩展一跳所得的子图密度, 用来反映中心基因所在区域的局部信息; 特征3: 中心基因的 N 个一阶邻居的 $|\log FC|$ 值, 表示为 $v_1, v_2, \dots, v_N$ ($v_1 \geq v_2 \geq \dots \geq v_N$), 用来反映中心基因一阶邻居的差异表达程度; 特征4和特征5分别表示中心基因 N 个一阶邻居 $|\log FC|$ 的均值和方差. 若中心基因不是癌症致病基因, 将向量的标签设置为0, 记作负样本; 若中心基因是癌症致病基因, 将向量的标签设置为1, 记作正样本. 上述5个特征均根据所归纳的两点特性而构造, 其中, 特征1用来刻画特性(i), 特征2和3用来刻画特性(ii), 特征4和5用来刻画邻域数据的分布.

获取基因节点的向量表示流程如图3所示. 图3A中, 对于21个用于构建训练集的癌症, 首先将该癌症的致病基因投影至背景网络作为中心基因, 按照如图2所示的结构, 计算表示该基因节点特征的向量, 组成当前癌症的正样本集; 接着随机从网络中选取非致病基因作为中心基因, 计算表示该基因节点特征的向量, 得到当前癌症的负样本集; 重复上述过程, 所有用于训练的癌症的正负样本集构成总的训练集. 图3B中, 对10个用于构建测试集的癌症, 将背景网络中的每个节点依次作为中心节点, 计算表示该基因节点特征的向量, 每个癌症得到一个规模为13397的测试集.

(2) SVM分类模型的性能. 使用小规模数据来训练支持向量机时, 若正负样本比例相差较大, 会导致预测结果产生不合理的偏移. 训练集中有227个正样本, 1185个负样本, 为了保证训练时样本比例均衡, 定义了如下的数据处理流程: (i) 从未使用过的负样本中

无放回地抽取227个样本; (ii) 将所有正样本和抽取的负样本作为输入, 从中分出20%作为验证集用于评估分类器性能; (iii) 重复步骤(i)和步骤(ii)5次得到5个模型, 将5个模型在各自验证集下的性能取均值得到SVM分类器模型的整体性能.

按照上述数据处理流程, 将分类器同参数独立训练100次, 将100次的性能取均值得到结果如图4所示, 图注释格式为: 归一化方法_参数 N . 参数 N 代表图2中为特征3设置的固定长度: 若中心基因的一阶邻居个数小于 N , 则末尾用0填补; 若一阶邻居个数大于 N , 则将每个邻居的 $|\log FC|$ 值由大到小排列, 取前 N 个作为 v_1 到 v_N . 若 $v_i (1 \leq i \leq N)$ 对应的一阶邻居在对应癌症的差异表达基因分析结果中的FDR大于0.5, 则将 v_i 置为0, 反之 v_i 保持不变. 当归一化方法的标识为no时, 代表不对 $|\log FC|$ 值进行处理; 若为max-min时, 则对 $|\log FC|$ 采用最大最小归一化; 若为z-score时, 则对 $|\log FC|$ 采用标准化分数归一化.

由图4可知, SVM分类模型最高的精度值和AUC值均出现在中心基因的一阶邻居个数参数(N)设置为300, 归一化方法设置为最大最小归一化(max-min)时. 此归一化方法可以尽可能地保持数据的原始分布, 因此, 保留该组参数下的SVM模型, 依次对表3中的10种癌症所对应的测试集进行类标签预测.

(3) 癌症致病基因的预测结果. 基于上文确定的分类模型, 对未知致病基因的癌症(测试集)进行类标签预测. 对每个癌症, 同参数独立预测100次, 将预测分数取均值, 若均值大于0.5, 则将中心基因判定为致病基因, 小于0.5判定为非致病基因. 每种癌症预测所得的致病基因数量如表3所示. 由表3可知, 测试集中每个癌症预测所得的致病基因数目约占当前癌症总预测基因数目(13397)的5%~13%.

2.4 癌症致病基因预测结果的验证

(1) ICGC 数据库验证. 国际癌症基因组联盟(ICGC)收录整理了来自全球50种具有临床和社会重

特征1	特征2	特征3	特征4	特征5
中心基因的 $ \log FC $ 值	子图密度	中心基因前 N 个一阶邻居($v_1, \dots, v_N$)的 $ \log FC $ 值	特征3的均值	特征3的方差

图2 基因节点特征的向量表示

Figure 2 Vector representation of gene node features

图 3 获取基因节点的向量表示

Figure 3 Extract vector representation of gene nodes

图 4 不同参数下SVM分类模型的性能对比

Figure 4 Performance comparison of SVM classification models under different parameters

表 3 SVM分类器模型预测结果

Table 3 Prediction results of the SVM classifier model

癌症名称	预测所得的致病基因个数	癌症名称	预测所得的致病基因个数
ACC	861	NB	611
CHOL	1341	ORCA	1190
MPM	682	OSTE	812
MPMD	972	PNET	1694
NACA	905	TNCA	1139

要性的不同癌症类别或亚型的大规模癌症基因组研究数据, 包括基因异常表达、体细胞突变、表观遗传修饰等^[35]. 本数据库关注基因变化给相应癌症带来的影响, 并将此影响划分为3个等级, 分别表示变异程度高的致病基因、变异程度低的致病基因以及未知变异程度的致病基因, 标识为高(high-impact)、低(low-impact)、未知(unknown-impact). 在10种用于预测的癌症中, 3种癌症在ICGC数据库中存在记录, 分别为胆管癌、鼻咽癌和口腔癌, 见表4. CHOL的预测结果中, 排名前20的致病基因均与该疾病相关, 其中10个具有高影响、3个低影响、7个影响程度未知. ORCA的预测结果中, 排名前20的致病基因也均与该疾病相关, 其中8个具有高影响、2个低影响、10个影响程度未知. NACA的预测结果中, 排名前20的致病基因中有5个与该疾病相关, 其中3个具有高影响、1个低影响、1个影响程度未知. 对于研究程度较高的疾病, 其相关记录丰富, 如CHOL. 但对于研究程度不高的疾病, 如NACA, 其数据来源有限是验证结果出现较大差异的主要原因.

(2) 结果对比验证及文献验证. 本节选择3个包含肿瘤相关基因的数据库COSMIC^[36], NCG^[37]和OncoKB^[38], 研究上文所述的10种癌症的致病基因之间

的共性与特性. 其中均收录了来自人类癌症基因普查^[39]等项目所标注的癌症基因(即在获得体细胞改变后在驱动癌症中起作用的基因). 将在该10种癌症预测结果内排名前20的致病基因中, 出现次数大于等于6次的基因作为高频致病基因, 将只与某一癌症相关的基因记为该癌症的特异性致病基因. 结果显示, 高频致病基因共14个, *SRC*, *GRB2*, *SPI*, *PRKCA*出现次数为10, *YWHAZ*, *TP53*, *CSNK2A1*, *EGFR*出现次数为9, *CDK1*出现次数为8, *JUN*, *MAPK1*, *PRKACA*出现次数为7, *FYN*, *ABL1*出现次数为6. 其中三个数据库均有验证的有6个, 至少有一个数据库有验证的为10个.

对于特异性致病基因, 预测结果显示ACC的特异性致病基因为*CDKN1A*和*PLK1*. Smith等人^[40]表明*PLK1*抑制剂有望作为治疗ACC患者亚型的候选药物. *TP53*种系突变存在于多达80%的Li-Fraumeni综合征(罕见癌症易发性疾病)中, *CDKN1A*作为编码p53在细胞周期停滞的主要介导基因, 其突变与该癌症的发生有着密切关系^[41].

CHOL的特异性致病基因为*ALB*, *APOA1*, *HNF4A*, *HNF1A*, *F2*, *ALDH2*, *ALDH9A1*以及*ALDH4A1*和*ALDH3A2*. 对于*APOA1*基因. Wang等人^[42]研究表明, 它的表征可作为CHOL的潜在生物标志. 对于4个ALDH族基因, 其过度表达与癌症干细胞样特征和不良的癌症预后有关, 且在癌症干细胞中ALDH具有高活性^[43]. 同时Chen等人^[44]也通过研究表明, *ALDH1A3*对CHOL患者预后和相关药物的耐药性存在影响.

MPM的特异性致病基因为*CSNK2A2*和*EP300*. 靶向CREBBP或EP300抑制蛋白可用于恶性胸膜间皮瘤的基因治疗和病毒治疗^[45], 且该文献所提及的*CREBBP*基因也出现在了MPM预测结果的排名前20的致病基因中. *CSNK2A2*是高度保守且普遍存在的丝

表 4 ICGC数据库验证结果

Table 4 ICGC database verification results

癌症名称	高影响基因	低影响基因	影响程度未知的基因
CHOL	<i>ALB</i> , <i>HNF4A</i> , <i>SPI</i> , <i>HNF1A</i> , <i>ABL1</i> , <i>F2</i> , <i>TP53</i> , <i>ALDH4A1</i> , <i>ALDH9A1</i> , <i>FNI</i>	<i>APOA1</i> , <i>PRKCA</i> , <i>ALDH2</i>	<i>GRB2</i> , <i>PRKACA</i> , <i>SRC</i> , <i>YWHAZ</i> , <i>CREB3</i> , <i>CDK1</i> , <i>ALDH3A2</i>
NACA	<i>KRT40</i> , <i>NCK1</i> , <i>TP53</i>	<i>SRC</i>	<i>GOLGA2</i>
ORCA	<i>SRC</i> , <i>APC</i> , <i>SMAD3</i> , <i>ACTB</i> , <i>YWHAZ</i> , <i>TP53</i> , <i>FNI</i> , <i>FOS</i>	<i>EGFR</i> , <i>MAPK1</i>	<i>GRB2</i> , <i>FYN</i> , <i>SPI</i> , <i>CSNK2A1</i> , <i>JUN</i> , <i>YWHAZ</i> , <i>CDK1</i> , <i>PIK3R1</i> , <i>PRKCA</i> , <i>CDK2</i>

氨酸或苏氨酸激酶, 在许多肿瘤中均过表达^[46].

MPMD的特异性致病基因为*CTNNB1*以及*ARRB2*. 免疫组织化学实验显示, 一种可能存在于胸膜中的局部侵袭性肿瘤通常具有*CTNNB1*基因突变^[47]. 对于*ARRB2*, 虽未明确找到其调控或诱发该疾病的相关文献, 但Kallifatidis等人^[48]表明*ARRB*参与调节膀胱癌的癌症干细胞, 并且是肿瘤化学反应疗法的潜在预后指标.

NB的特异性致病基因为*APP*和*PLCG1*. LeBlanc等人^[49]研究发现, 由于*APP*基因表达被剥夺, 细胞增殖降低了30%, 表明*APP*可能参与了神经突的延伸, 与成神经细胞瘤有潜在联系. 且在神经母细胞瘤的显著富集通路上存在*PLCG1*和*PLCG2*^[50].

ORCA的特异性致病基因为*FOS*. Mishra等人^[51]研究表明, AP-1超家族蛋白的差异表达、激活或病变的严重程度在人类口腔癌发生的过程中有着关键作用. *FOS*基因, 以及同样出现在ORCA预测结果的排名前20致病基因中的*JUN*都属于AP-1超家族.

NACA的特异性致病基因为*KRTAP10-3*, *NOTCH2NLA*和*MDF1*. 其中, *NOTCH2NLA*基因的表达对喉癌发挥了一定的作用^[52]. *KRTAP10-3*和*MDF1*未明确找到与NACA相关的证明. 但Sui等人^[53]表明, *MDF1*和*MDFIC*在结直肠癌中具有肿瘤调节功能.

OSTE的特异性致病基因为*ITGB1*, *MAPK14*和*PRKCD*. 近年GeneXplain研究团队^[54]通过网络分析等方法提出了*PRKCD*和*DUSP5*是有望治疗OSTE的药物靶标. *MAPK14*作为多种生化事件的整合点被选取作为治疗OSTE的核心靶点^[55]. *ITGB1*是表达最高的整合素蛋白, 通常在人类恶性肿瘤中过表达, 它可以通过Hedgehog信号通路, 增强OSTE的增殖和侵袭^[56].

PNET的特异性致病基因有4个. 其中*SMAD2*会引起神经外胚层基因的表达水平发生改变^[57]. *CALMI*是新型神经内分泌特异性肿瘤基因的鉴定过程中具有显著差异的基因^[58]. Chen等人^[59]使用免疫组织化学法检测肿瘤中*VIM*的表达来指导对原发性PNET的预后因素分析. *EEF1A1*调节细胞骨架, 它促进肿瘤发生, 阻断细胞凋亡并增加病毒发病^[60].

TNCA的特异性致病基因为*LCK*和*LYN*. *LCK*在T细胞信号转导中起关键作用, 最终该信号将影响如*PI-3K*等靶标, 且*PI-3K*的过度活性会促进不受控制的细胞生长, 因此*LCK*可作为口腔扁平苔藓和口腔鳞状细

胞癌的分子生物标志物^[61]. 在口腔鳞状细胞癌年轻患者的预后标志物的研究中, Gu等人^[62]表明LYN癌症驱动基因存在体细胞变异.

(3) 富集分析及相关通路验证. 为了弥补单个基因单个验证的局限性, 本节将预测的所有潜在致病基因记为注释基因集, 分别进行KEGG(Kyoto encyclopedia of genes and genomes)和DO(disease ontology)富集分析, 从基因功能和信号通路方面对预测结果进行宏观的解释.

KEGG是系统分析基因功能的知识基础, 存储了基因参与人体各通路的信息^[29]. 富集参数设置p.adjust的阈值为0.1, 所得富集结果如图5A所示. 图中的横坐标表示富集在相应通路上的基因所占注释基因集的比例. 图中展示了置信度最高的30条通路, 观察可知, 进行富集的注释基因集与癌症中蛋白聚糖、病毒致癌、胰腺癌、FoxO信号通路、MAPK信号通路、ErbB信号通路、前列腺癌、PI3K-Akt信号通路和非小细胞肺癌等9条通路, 以及许多免疫相关的信号通路都有着密切联系. FoxO是叉头转录因子家族的一个亚族, 决定细胞命运且在多种癌症中有着重要的抑癌作用, 与FoxO信号通路相互作用的最主要的通路为PI3K-Akt^[63]. 相关研究表明, PI3K-Akt信号通路是参与细胞生长、增殖、代谢、运动和凋亡等细胞过程的关键调节因子, 该通路的异常激活促进了许多人类癌症中肿瘤细胞的存活和增殖^[64]. 同样, MAPK信号通路和ErbB信号通路也是癌症的产生和发展进程中发挥重要作用的信号传导通路^[65,66], 可用来研究癌症的药物敏感性和耐药性等.

疾病本体论(DO)数据库是收录了8043种遗传、发育和获得性人类疾病的综合知识库, 主要被用于疾病注释^[67]. 富集参数设置p.adjust的阈值为0.1, 所得富集结果如图5B所示. 图中的横坐标表示富集在相应疾病上的基因所占注释基因集的比例. 图中展示了置信度最高的30个富集疾病, 观察可知, 与注释基因集密切相关的疾病大多为癌症, 在排名前30中占有26个, 例如, 乳腺癌、肌肉骨骼系统癌、结缔组织癌、前列腺癌、骨肉瘤、结直肠癌、甲状腺癌、卵巢癌等. 由于较小的注释基因集会导致富集结果产生较大的偏差, 本文并未对单个验证癌症的预测结果来进行富集分析, 仅从相对宏观的角度说明了该基因节点特征的向量表示方法的可行性.

图 5 KEGG(A)和DO(B)富集分析结果可视化

Figure 5 Visualization of KEGG (A) and DO (B) enrichment analysis results

3 讨论

本文从观察到的生物学现象出发, 通过已有的数据和可参考的金标准来总结癌症致病基因的相关特性, 并基于该特性构造基因节点特征的向量表示方法, 最终通过完成癌症致病基因的预测任务来说明前文所述特性的合理程度. 本节将从分类模型和评价指标的选择以及验证结果两方面进一步讨论, 更充分地阐明该特性对癌症致病基因与全基因组中的其他基因具有较好的区分能力.

在进行分类模型和评价指标选择的过程中, 本文坚持采用尽可能简便的模型和更适用于当前任务的评价方法来说明假设的合理性. 支持向量机是一种传统、简便, 又运用广泛的二分类预测模型. 在2.3节癌症致病基因预测的实验中, SVM模型在通过前文所述特性所提取的训练集上取得了相对可观的性能, Precision值可达0.950, AUC值可达0.808, 但仍存在可提升空间. 在尝试采用了一些常见集成分类器后, 分类性能有一定提升, 却由于样本量不足, 效果提升不够明显, 且容易产生过拟合现象, 同时庞大繁琐的分类器反而会遮盖样本本身的差异, 无法突出特征在完成癌症致病预测任务中的主体地位. 因此, 文中仅对SVM模型的预测结果进行展示说明. 除Precision值以

外, 常见的评价分类性能的指标还有准确率和召回率等, 鉴于该任务更关注于预测致病基因而非非致病基因的能力, 所以选择评价指标的重心为预测出的致病基因是否为真正的致病基因, 即Precision值所代表的含义. 相异于专门构造机器学习模型来进行致病基因识别的方法, 使用这样精简的现有模型可以在最大程度上突出假设本身的特性, 更有力地表明本文所总结的癌症致病基因在人类蛋白质相互作用网络上的特征.

分析2.4节的验证结果可得, 对于广泛存在于各个验证癌症中的高频致病基因, 该类基因的扰动与癌症这一类复杂疾病可能存在密切关系, 它们同时出现在了绝大多数验证癌症的排名前20预测结果中, 其引起的表型变化可反映不同癌症症状之间的共同特征, 因此, 挖掘此类基因将有助于对新型癌症的初步认识和初步治疗. 对于仅存在于某个癌症中的特异性致病基因, 该类基因与特定癌症可能存在特定表型相互关联, 若能准确识别此类基因将大大推动癌症精准诊疗的发展.

未来的研究工作将从癌症致病基因的预测任务迁移至更多疾病致病基因的预测任务, 探究本文所述的特征构造思想的普适性, 寻求其他生物学网络的信息和患病前后基因表达的变化信息之间更深层的联系.

参考文献

- Cheng L, Yang H, Zhao H, et al. MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Briefings BioInf*, 2019, 20: 203–209
- Ma H, Marti-Gutierrez N, Park S W, et al. Correction of a pathogenic gene mutation in human embryos. *Nature*, 2017, 548: 413–419
- Pottinger T D, Puckelwartz M J, Pesce L L, et al. Pathogenic and uncertain genetic variants have clinical cardiac correlates in diverse biobank participants. *J Am Heart Assoc*, 2020, 9: 26
- Zou Y, Hui R, Song L. The era of clinical application of gene diagnosis in cardiovascular diseases is coming. *Chronic Dis Transl Med*, 2019, 5: 214–220
- Bu L L, Yan J, Wang Z, et al. Advances in drug delivery for post-surgical cancer treatment. *Biomaterials*, 2019, 219: 119182
- Cohen J D, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 2018, 359: 926–930
- Zhang X, Zou Q, Rodriguez-Paton A, et al. Meta-path methods for prioritizing candidate disease miRNAs. *IEEE ACM Trans Comput Biol Bioinf*, 2019, 16: 283–291
- Zhang H P, Wang H N, Lu G M, et al. Finding differentially co-expressed disease-related genes based on mutual information (in Chinese). *J Southeast Univ (Nat Sci Ed)*, 2009, 39: 151–155 [张焕萍, 王惠南, 卢光明, 等. 基于互信息的差异共表达致病基因挖掘方法. 东南大学学报 (自然科学版), 2009, 39: 151–155]
- Adie E, Adams R R, Evans K L, et al. Speeding disease gene discovery with SUSPECTS. *BMC BioInf*, 2005, 6: 1–3
- Yu S, Van Vooren S, Tranchevent L C, et al. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text

- mining. *Bioinformatics*, 2008, 24: i119–i125
- 11 Greenman C, Wooster R, Futreal P A, et al. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 2006, 173: 2187–2198
- 12 Momozawa Y, Iwasaki Y, Parsons M T, et al. Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nat Commun*, 2018, 9: 4083
- 13 Belot A, Rice G I, Omarjee S O, et al. Contribution of rare and predicted pathogenic gene variants to childhood-onset lupus: a large, genetic panel analysis of British and French cohorts. *Lancet Rheumatol*, 2020, 2: e99–e109
- 14 Li X, Chang M, Wang L. Information recognition of pathogenic modules in gene statistics of big data. *Nanomater Energy*, 2021, 10: 35–42
- 15 Collier O, Stoven V, Vert J P. LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput Biol*, 2019, 15: e1007381
- 16 Luo P, Ding Y, Lei X, et al. deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front Genet*, 2019, 10: 12
- 17 Wang C Y, Zhang J, Wang X P, et al. Pathogenic gene prediction algorithm based on heterogeneous information fusion. *Front Genet*, 2020, 11: 5
- 18 Timilsina M, Yang H, Sahay R, et al. Predicting links between tumor samples and genes using 2-layered graph based diffusion approach. *BMC Bioinf*, 2019, 20: 1–20
- 19 Xu B, Liu Y, Yu S, et al. A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. *BMC Med Genomics*, 2019, 12: 188
- 20 Liu X, Tang W H, Zhao X M, et al. A network approach to predict pathogenic genes for *Fusarium graminearum*. *PLoS ONE*, 2010, 5: e13021
- 21 Weinstein J N, Collisson E A, Mills G B, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 2013, 45: 1113–1120
- 22 Amberger J S, Bocchini C A, Schiettecatte F, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*, 2015, 43: D789–D798
- 23 Barrett T, Troup D B, Wilhite S E, et al. NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Res*, 2007, 35: D760–D765
- 24 Aranda B, Achuthan P, Alam-Faruque Y, et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 2010, 38: D525–D531
- 25 Ceol A, Chatr Aryamontri A, Licata L, et al. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 2010, 38: D532–D539
- 26 Stark C, Breitkreutz B J, Chatr-Aryamontri A, et al. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*, 2011, 39: D698–D704
- 27 Prasad T S K, Goel R, Kandasamy K, et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res*, 2009, 37: D767–D772
- 28 Ruepp A, Waegele B, Lechner M, et al. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res*, 2010, 38: D497–D501
- 29 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 2000, 28: 27–30
- 30 Lee D S, Park J, Kay K A, et al. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA*, 2008, 105: 9880–9885
- 31 Matys V, Fricke E, Geffers R, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 2003, 31: 374–378
- 32 Hornbeck P V, Kornhauser J M, Tkachev S, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res*, 2012, 40: D261–D270
- 33 Wong A K, Krishnan A, Troyanskaya O G. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic Acids Res*, 2018, 46: W65–W70
- 34 Osuna E, Freund R, Girosi F. Support Vector Machines: Training and Applications. Cambridge, MA: Massachusetts Institute of Technology, 1998
- 35 Hudson T J, Anderson W, Artez A, et al. International network of cancer genome projects. *Nature*, 2010, 464: 993–998
- 36 Tate J G, Bamford S, Jubb H C, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, 2019, 47: D941–D947
- 37 Repana D, Nulsen J, Dressler L, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol*, 2019, 20: 12
- 38 Chakravarty D, Gao J, Phillips S M, et al. OncoKB: a precision oncology knowledge base. *JCO Precision Oncol*, 2017, 1: 1–16
- 39 Futreal P A, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer*, 2004, 4: 177–183
- 40 Smith G, Liang R, Chortis V, et al. SUN-LB22 PLK1 as a new treatment target for adrenocortical carcinoma. *J Endocrine Soc*, 2020, 4: LB22
- 41 Andrade R C, Dos Santos A C E, de Aguiar Neto J C, et al. TP53 and CDKN1A mutation analysis in families with Li-Fraumeni and Li-Fraumeni like syndromes. *Familial Cancer*, 2017, 16: 243–248

- 42 Wang X, Dai S, Zhang Z, et al. Characterization of apolipoprotein A-I as a potential biomarker for cholangiocarcinoma. *Eur J Cancer Care*, 2009, 18: 625–635
- 43 Chang P M H, Chen C H, Yeh C C, et al. Transcriptome analysis and prognosis of ALDH isoforms in human cancer. *Sci Rep*, 2018, 8: 2713
- 44 Chen M H, Weng J J, Cheng C T, et al. ALDH1A3, the major aldehyde dehydrogenase isoform in human cholangiocarcinoma cells, affects prognosis and gemcitabine resistance in cholangiocarcinoma patients. *Clin Cancer Res*, 2016, 22: 4225–4235
- 45 Fukazawa T, Matsuoka J, Naomoto Y, et al. Malignant pleural mesothelioma-targeted CREBBP/EP300 inhibitory protein 1 promoter system for gene therapy and virotherapy. *Cancer Res*, 2008, 68: 7120–7129
- 46 Wang G, Ahmad K A, Harris N H, et al. Impact of protein kinase CK2 on inhibitor of apoptosis proteins in prostate cancer cells. *Mol Cell Biochem*, 2008, 316: 91–97
- 47 Galateau-Salle F, Churg A, Roggli V, et al. The 2015 world health organization classification of tumors of the pleura: advances since the 2004 classification. *J Thoracic Oncol*, 2016, 11: 142–154
- 48 Kallifatidis G, Smith D K, Morera D S, et al. β -arrestins regulate stem cell-like phenotype and response to chemotherapy in bladder cancer. *Mol Cancer Ther*, 2019, 18: 801–811
- 49 LeBlanc A C, Kovacs D M, Chen H Y, et al. Role of amyloid precursor protein (APP): study with antisense transfection of human neuroblastoma cells. *J Neurosci Res*, 1992, 31: 635–645
- 50 Petrov I, Suntsova M, Ilnitskaya E, et al. Gene expression and molecular pathway activation signatures of *MYCN*-amplified neuroblastomas. *Oncotarget*, 2017, 8: 83768–83780
- 51 Mishra A, Bharti A C, Saluja D, et al. Transactivation and expression patterns of Jun and Fos/AP-1 super-family proteins in human oral cancer. *Int J Cancer*, 2009, 126: 819–829
- 52 Cabezas-Camarero S, García-Barberán V, Hernánpérez-Hidalgo D, et al. Mutational profile of dysplastic lesions evolving to laryngeal cancer. *J Clin Oncol*, 2019, 37: e17551
- 53 Sui Y, Li X, Oh S, et al. Opposite roles of the JMJD1A interaction partners MDFI and MDFIC in colorectal cancer. *Sci Rep*, 2020, 10: 8710
- 54 Kolpakov F, Porokov V, Selivanova G, et al. GeneXplain-identification of causal biomarkers and drug targets in personalized cancer pathways. *J Biomol Tech JBT*, 2011, 22(Suppl): 16
- 55 Tan J, Wu Z, Chen J, et al. Bioinformatic characteristics and therapeutic mechanisms of calycosin-anti-bisphenol a-related osteosarcoma. Research Square, 2020, <https://doi.org/10.21203/rs.3.rs-77018/v1>
- 56 Yao Z H, Han L, Chen Y B, et al. Hedgehog signalling in the tumorigenesis and metastasis of osteosarcoma, and its potential value in the clinical therapy of osteosarcoma. *Cell Death Dis*, 2018, 9: 701
- 57 Sakaki-Yumoto M, Liu J, Ramalho-Santos M, et al. Smad2 is essential for maintenance of the human and mouse primed pluripotent stem cell state. *J Biol Chem*, 2013, 288: 18546–18560
- 58 Hofslie E, Wheeler T E, Langaas M, et al. Identification of novel neuroendocrine-specific tumour genes. *Br J Cancer*, 2008, 99: 1330–1339
- 59 Chen L J, Jia Y X, Fan F F, et al. Expression of FLI-1 and analysis of prognostic factors in primitive neuroectodermal tumor (in Chinese). *Chin J Oncol*, 2010, 32: 917–920 [陈利娟, 贾永旭, 范菲菲, 等. 原始神经外胚层肿瘤中FLI-1的表达及预后因素分析. 中华肿瘤杂志, 2010, 32: 917–920]
- 60 Abbas W, Kumar A, Herbein G. The eEF1A proteins: at the crossroads of oncogenesis, apoptosis, and viral infections. *Front Oncol*, 2015, 5: 75
- 61 Oluwadara O, Giacomelli L, Christensen R, et al. LCK, survivin and PI-3K in the molecular biomarker profiling of oral lichen planus and oral squamous cell carcinoma. *Bioinformation*, 2009, 4: 248–257
- 62 Gu X, Coates P J, Boldrup L, et al. Copy number variation: a prognostic marker for young patients with squamous cell carcinoma of the oral tongue. *J Oral Pathol Med*, 2019, 48: 24–30
- 63 Farhan M, Wang H, Gaur U, et al. FOXO signaling pathways as therapeutic targets in cancer. *Int J Biol Sci*, 2017, 13: 815–827
- 64 Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR signaling in cancer. *Front Oncol*, 2014, 4: 64
- 65 Lee S, Rauch J, Kolch W. Targeting MAPK signaling in cancer: mechanisms of drug resistance and sensitivity. *Int J Mol Sci*, 2020, 21: 1102
- 66 Hynes N E, MacDonald G. ErbB receptors and signaling pathways in cancer. *Curr Opin Cell Biol*, 2009, 21: 177–184
- 67 Schriml L M, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, 2012, 40: D940–D946

Prediction of cancerous pathogenic genes based on network and gene differential expression information

YU Liang & REN ShuJie

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

In genome research, the discovery of disease-causing pathogenic genes is a major challenge in genome research. Recently, using the available biological data, many researchers have employed computational methods for pathogenic gene prediction. However, most of these computational methods are based on gene interaction networks or other similar networks, and the potential connection between the local network of specific genes and their differential expression information is rarely considered. This paper explores the biological properties of cancerous pathogenic genes and their neighbors based on a local network structure and the differential expression information of genes. Furthermore, machine learning methods are employed for predicting cancerous pathogenic genes based on the newly discovered biological properties. First, the expression data of 21 cancers and their pathogenic genes are obtained from the TCGA (The Cancer Genome Atlas) and OMIM (Online Mendelian Inheritance in Man) databases. Human-protein interaction and tissue-specific interaction networks corresponding to each cancer gene are used as background networks to analyze the potential biological properties between the neighborhood information of different networks and the changed information of the gene expression. Then, based on the discovered biological properties, a vector representation method for gene node features is established, and the support vector machine is employed for pathogenic gene prediction. Experimental results are verified using ICGC (International Cancer Genome Consortium), COSMIC (Catalogue Of Somatic Mutations In Cancer), NCG (Network of Cancer Genes), OncoKB (Oncology Knowledge Base), and other standard databases, as well as using related literature, disease annotations and pathway enrichments. Results show that the definition of gene features can be used to distinguish cancerous pathogenic genes from other genes, thus providing powerful hypotheses for cancerous pathogenic gene prediction. Furthermore, reliable pathogenic gene candidates are obtained for related biological experiments to promote research on the pathogenic mechanism of cancer, a complex disease.

cancerous pathogenic genes, interaction network, differential expression information, vector representation of features, support vector machine

doi: [10.1360/SSV-2021-0324](https://doi.org/10.1360/SSV-2021-0324)