基于 Transformer 特征学习与多源异构数据融合的泛癌驱动模块识别

邓诗宇
广西工业职业技术学院　广西壮族自治区南宁市　530000

【摘　要】针对传统方法处理高维稀疏数据能力不足的问题，本文提出基于Transformer的深度特征学习框架。该框架以TCGA泛癌数据为基础，融合基因组、转录组及蛋白质互作网络等多源异构数据，利用多头自注意力机制捕捉基因间的非线性与长距离依赖关系，实现了对潜在驱动基因的深度语义表征，为多组学数据融合与癌症分子机制解析提供了新思路。

【关键词】癌症驱动模块；Transformer；多源异构数据；泛癌

A Pan-Cancer Driver Module Based on Transformer Feature Learning and Multi-Source Heterogeneous Data Fusion

Deng Shiyu
Guangxi Polytechnic University, Nanning 530000 Guangxi Zhuang Autonomous Region.

【Abstract】To address the limitations of traditional methods in processing high-dimensional sparse data, this paper proposes a deep feature learning framework based on Transformers. Built upon TCGA pan-cancer data, the framework integrates multi-source heterogeneous data including genomic, transcriptomic, and protein interaction networks. By employing a multi-head self-attention mechanism to capture nonlinear and long-range dependencies between genes, it achieves deep semantic representations of potential driver genes, providing novel insights for multi-omics data integration and the elucidation of cancer molecular mechanisms.

Keywords：Cancer-driven module; Transformer; Multi-source heterogeneous data; Pan-cancer

1 引言

癌症是基因组在多维度、多层次上协同失控的系统性疾病。传统方法将多组学数据视为孤立线性叠加，忽略了基因变异通过调控网络级联放大为恶性表型的机制。为此，本研究致力于构建深度融合多源异构数据的计算框架，通过先进的特征学习技术，从系统高度解析基因深层互作逻辑，精准锚定驱动癌症的核心功能模块。

2 数据

为了系统性地验证模型的有效性与鲁棒性，本研究基于真实癌症数据集进行实验测试。TCGA 泛癌体细胞突变数据包括 3110 个癌症样本和 11 种癌症类型的 11565 个基因。PPI 网络使用的是 HINT+HI2012[1–3]，并且保留了在体细胞突变数据中同时存在的基因，处理后的数据如下：癌症样本数 $m = 3110$，基因数 $n = 6930$，边数 $|E| = 25251$。通过 mirDIP 数据库[4] 可以得到基因–microRNA 网络，该网络由 6145 个基因、2734 个 microRNA 以及基因与 microRNA 的 229135 条边组成。

3 定义和符号

给定一组癌症样本 $S = \{s_i \mid i = 1,2,\ldots,m\}$ 及突变基因 $G = \{g_j \mid j = 1,2,\ldots,n\}$，定义体细胞突变矩阵 $A_{m\times n}$、连通的 PPI 网络 $PP = (V,E)$ 以及基因–microRNA 网络 $PM = (V^g, V^m, E^{gm}, W^{gm})$，并引入互斥度 $MEX(M)$、覆盖度 $COV(M)$ 及模块相对大小 $RS(M_i)$ 等指标，用于量化驱动模块的互斥性与覆盖性。在此基础上，构建互斥度得分 $MS(P)$ 与覆盖度得分 $CS(P)$，进而定义驱动模块集得分 $\mathrm{D}MSS(P) = MS(P) \times CS(P)$。最终，将癌症驱动模块识别问题形式化为一个优化问题：在给定 PPI 网络 PP、体细胞突变矩阵 $A_{m\times n}$、基因–microRNA 网络 PM、基因规模总数 $Totalg$ 及模块最小规模 $Mins$ 等约束条件下，寻找一组不重叠的连通模块 $P = \{M_1, M_2, \ldots, M_r\}$，使得 $DMSS(P)$ 最大化。

4 方法

本节提出了一种基于 Transformer 特征学习的癌症驱动模块识别方法，该方法分为四个主要步骤：权重分配、特征

提取、聚类和构建驱动模块，如图 1 所示。

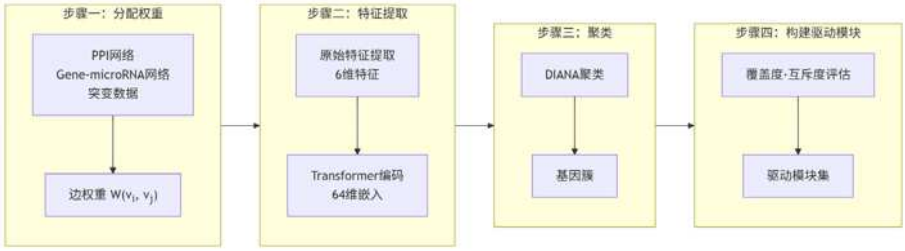


图 1　方法的步骤

首先，在权重分配步骤中，引入基因–microRNA 网络对 PPI 网络的边进行加权。给定基因对 g_i 和 g_j，首先基于其共同 microRNA 邻居计算置信度 $CF(g_i, g_j)$，再结合互斥度 $ME(g_i, g_j)$ 与覆盖度 $cov(\{g_j\})$，定义边权重 $w(v_i, v_j)$。当互斥度低于阈值 θ 时，权重设为 0。

其次，在特征提取步骤中，采用 Transformer 编码器对加权 PPI 网络中的节点进行特征提取。Transformer 是一种基于自注意力机制的深度学习模型，由 Vaswani 等人在 2017 年提出[6]。给定加权 PPI 网络 $G = (V, E, W)$，其中包含 6930 个基因节点，首先为每个节点提取 6 维原始拓扑特征（包括度、加权度、聚类系数、PageRank 值、邻居平均权重及特征向量中心性），构成特征矩阵 $X \in \mathbb{R}^{6930\times6}$。随后，对特征矩阵进行标准化处理，使其服从标准正态分布。Transformer 编码器通过线性投影将 6 维特征映射至 64 维模型空间，并添加位置编码以提供节点间的相对位置参考。编码器由 3 个相同的层堆叠而成，每层包含多头自注意力机制（4 个注意力头，每个头维度为 16）和前馈神经网络（隐藏层维度 256），并应用残差连接与层归一化。模型采用自监督预训练策略，通过重建原始特征图进行学习，损失函数为均方误差，使用 Adam 优化器迭代训练 100 轮。训练完成后，提取编码器最后一层的输出作为节点的最终 64 维特征表示 $H \in \mathbb{R}^{6930\times64}$。图 2 展示了 Transformer 特征提取的完整流程。

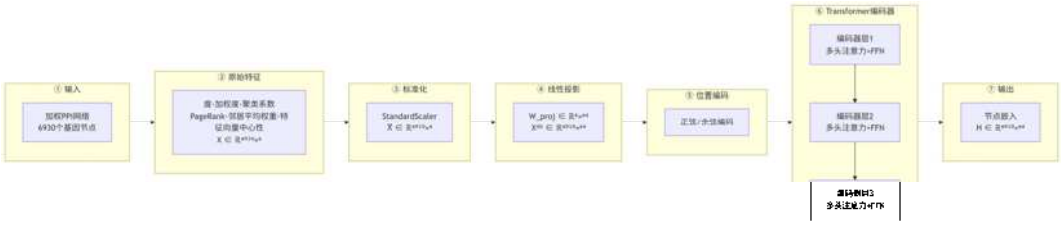


图 2　特征提取流程

在特征提取完成后，利用 DIANA 层次聚类算法对基因节点进行聚类。该算法采用马氏距离衡量节点特征向量间的相似度，通过迭代分裂的方式将初始的单一基因簇逐步划分为指定数量的子簇，从而得到一组基因簇集。

最后，基于聚类结果设计了一种启发式算法来构建最终的驱动模块。该算法通过迭代删除节点影响力最低的节点，并从每个基因簇中提取满足最小规模要求的连通分量，直至所有模块的基因总数不超过预设的规模上限，最终生成一组不重叠、内部连通且具有高互斥性与高覆盖度的候选癌症驱动模块。Transformer 方法通过自注意力机制捕捉全局依赖、多头注意力多角度挖掘语义特征，以及无需标注数据的自监督预训练，有效适用于此类无监督的驱动模块识别任务。

5 实验结果与分析

本节以 COSMIC 癌症基因普查（CGC）数据库[9] 和癌症基因网络（NCG）[10] 为参考基因集进行评估。表 1 展示了该方法在 *Totalg* = 100 时检测到的 12 个模块，模块大小在 3 到 24 之间。为了便于后续讨论，在表 1 中被 COSMIC 数据库覆盖的基因用加粗文字进行标注，当 *Totalg* = 100 时，共 60 个基因在 COSMIC 数据库中，共 59 个基因在癌症基因网络（NCG）中，每个基因模块均以其内部突变频率最高的基因进行命名。本研究参考 KEGG 数据库[8]，发现绝大多数识别出的模块均属于已知的致癌相关通路。

表 1　*Totalg* = 100 已识别的模块

模块	基因
PIK3CA	PIK3CA PIK3R1 CTNNB1 IGF1R CCND1 CDH1 MDM2 MDM4 EGFR TLN1 ATM IKBKB CDK4 VHL NEDD9 CDK6 ERBB2 CDKN2A IRS1 AKT1 RB1
MYC	CCNE1 MYC FBXW7

KAT6A	KAT6A NFE2L2 CREBBP
TP53	XPO1 ERCC6 TP53 NPM1 PTEN ING1
ABCA1	ARHGEF11 ARHGEF12 ABCA1
SETD2	NOTCH1 SETD2 ATN1 KIAA2026 LTBP1 SPEN BRCA1 ANKHD1 HIVEP1 CACNA1A ACACA RERE ATXN1
EP300	MEF2A KLF5 FLNA KDM6A REL NCOA3 EP300 NCOA6 ESR1 AR TCF3
FGFR3	PTK2 PTPN11 ITSN1 ITSN2 PTPRB PTPRC BRAF GTF3C1 PTPRK PTPRO EZH2 KIAA1549 INSR MYO5B HRAS FGFR3 MLLT4 MAP3K1 NEDD4 RASA1 ERBB4 KIT SOS1 SOS2
ARID1A	CHD4 ARID1A STK11 SMARCA2 SMARCA4
SMAD4	UHRF2 SMAD4 TFE3
CHD6	RBBP6 CHD6 ZFYVE9
STAG2	STAG2 SMC1A STAG1

PIK3CA 模块显著富集于 FoxO、PI3K-Akt 及 p53 信号通路，涉及 IGF1R、EGFR、AKT1、CDK4 等关键基因。其中，FoxO 在肿瘤调控中发挥重要作用，PI3K-Akt 既驱动肿瘤发展也影响治疗响应。MYC 模块主要参与 PI3K-Akt 通路（CCNE1、MYC）；TP53 模块涉及 p53 及鞘脂通路（TP53、PTEN）；EP300 模块富集于甲状腺激素通路（ESR1、NCOA3、EP300）。FGFR3 模块广泛参与 Ras、MAPK 及 PI3K-Akt 通路，涉及 HRAS、BRAF、KIT、SOS1 等基因。Ras 通路失调在恶性肿瘤中极为常见，MAPK 通路则协同激活 CREB，从而上调胶质母细胞瘤中 Cyclin D1 的表达。

6 结论

本文提出基于 Transformer 的泛癌驱动模块识别方法，其多头自注意力机制实现全局依赖建模，自监督预训练无需标注数据，多注意力头从多语义视角增强模块可分离性。结合 DIANA 聚类与启发式算法，从 11 种泛癌数据中识别出 12 个高可信驱动模块，显著富集于 PI3K-Akt、MAPK 等经典通路，并在 COSMIC 及 NCG 数据库中具有高覆盖度。本研究为多组学融合与癌症机制解析提供了新思路。

参考文献

[1]Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease[J]. BMC systems biology, 2012, 6(1): 1-12.

[2]Leiserson M D M, Vandin F, Wu H T, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes[J]. Nature genetics, 2015, 47(2): 106-114.

[3]Yu H, Tardivo L, Tam S, et al. Next-generation sequencing to generate interactome datasets[J]. Nature methods, 2011, 8(6): 478-480.

[4]Tokar T, Pastrello C, Rossos A E M, et al. mirDIP 4.1—integrative database of human microRNA target predictions[J]. Nucleic acids research, 2018, 46(D1): D360-D370.

[5]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998-6008.

[6]McLachlan G J. Mahalanobis distance[J]. Resonance, 1999, 4(6): 20-26.

[7]Kanehisa, Minoru, and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30.

[8]John G Tate, Sally Bamford, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer.[J]. Nucleic Acids Research, 2019, 47(D1): D941-D947.

[9]Dressler L, Bortolomeazzi M, Keddar M R, et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource[J]. Genome Biol, 2022, 23(1): 35.

作者简介：邓诗宇(1996-12)，男，瑶族，桂林市人，硕士，工程师，研究方向：数据挖掘、生物信息学；

课题名称：基于 Transformer 和多源异构数据的泛癌驱动模块识别研究，课题编号：GYKY2025021B。