一种基于多组学生物网络的癌症关键模块挖掘方法

郭茂祖¹, 武雪剑², 赵宁³, 刘晓燕², 王春宇^{2*}

1. 北京建筑大学电气与信息工程学院, 北京 100044

2. 哈尔滨工业大学计算机科学与技术学院, 哈尔滨 150000

3. 哈尔滨工业大学生命科学与技术学院, 哈尔滨 150000

* 通信作者. E-mail: chunyu@hit.edu.cn

收稿日期: 2017-05-16; 接受日期: 2017-06-12; 网络出版日期: 2017-11-14

国家自然科学基金 (批准号: 61532014, 61571163, 61402132, 61671189) 资助项目

摘要 癌症的产生与发展与人类生存环境、个体遗传因素等都存在密切关系, 鉴于癌症对人类生存与健康所造成的生命危险, 全世界很多顶级科研机构都在积极开展研究, 以期尽早发现癌症发生、发展的致病机理. 新一代测序技术的出现极大地加速了癌症研究, 为发现癌症相关的重要信息奠定了基础. 本文从数据集成的角度研究异构多层次数据, 科学地引入 lncRNA (long non-coding RNA) 组学数据, 构建生物网络模型. 通过聚类方法挖掘致癌基因模块. 总结出一种挖掘导致癌症发生的关键基因模块的系统方法, 挖掘出包含 15 个基因的关键基因模块. 通过分析这些基因的功能与通路, 发现其与癌症密切相关. 通过生存分析, 发现它们能够很好地区分高低风险组. 所有这些结果表明通过集成多组学生物数据能够发现关键基因模块及其异常调控的基因集合, 有助于癌症研究.

关键词 多组学数据, 生物网络模型, 密度聚类, lncRNA, 异常调控

1 引言

随着高通量技术的飞速发展, 我们可以准确地获取生物体发育过程中各个阶段的各种生物数据^[1], 这些数据统称为多组学数据 (multi-omics data), 包括基因表达、基因调控、蛋白质/RNA 交互、变异数据、甲基化数据等. 这些数据对癌症的诊断与治疗提供了数据分析的基础. 为基于数据驱动研究癌症基因模式提供了有力的渠道.

近几年, 科研工作者利用多组学数据进行了理论研究, 并取得一定的研究进展. Ideker 等^[2] 基于网络方法, 通过集成基因表达与蛋白质交互数据预测乳腺癌转移, 并显著提升预测准确性. 中国科学院 Zhang 等^[3] 通过集成基因表达、DNA 甲基化与 microRNA 数据构建多重异构网络, 进而挖掘多

引用格式: 郭茂祖, 武雪剑, 赵宁, 等. 一种基于多组学生物网络的癌症关键模块挖掘方法. 中国科学: 信息科学, 2017, 47: 1510–1522, doi: 10.1360/N112017-00108
Guo M Z, Wu X J, Zhao N, et al. A method for mining core modules of cancer based on multi-omics biological network (in Chinese). Sci Sin Inform, 2017, 47: 1510–1522, doi: 10.1360/N112017-00108

维模块,分析模块的基因调节关系. Hoadley 等^[4]集成基因组和表观组数据研究泛癌症,揭示不同生物数据对癌症分类的影响. Yang 等^[5]利用 TCGA 数据,集成 4 种癌症的基因共表达网络,识别预后 mRNA 和 microRNA 基因. 研究显示在癌症基因共表达网络中预后 mRNA 和 microRNA 的目标基因具有相似的特征. Han 等^[6]集成了来自 TCGA 的 7 个癌症的 2808 个病人样本 RNA-seq 数据构建假基因表达谱,识别了差异表达的假基因,并且对癌症亚型进行分类,预测一些癌症可能的临床亚型. Li 等^[7]整合了多组学数据包括基因表达与 miRNA 表达、拷贝数、甲基化、转录因子,提出回归分析方法发现急性髓细胞样白血病的调控因子. Ping 等^[8]给出了识别核心基因模块及失调基因的系统化方法,通过集成多维基因组数据,构建多层次因素网络模型,将该方法应用于恶性胶质瘤. 以上数据集成方法针对不同类型复杂疾病关联模式的研究,集成了不同层面的生物数据.

近 50 年来许多国家都报道肺癌的发病率和死亡率均明显增高^[9]. 肺癌的病因至今尚不完全明确. 在肺癌的各种类型中,非小细胞肺癌占肺癌的 80% 左右,其中肺鳞癌 (lung squamous cell carcinoma, LUSC) 占肺癌的 50% 左右. 目前,国内外的许多学者都在研究肺癌. 长链非编码 RNA (lncRNA) 是长度大于 200 nt 的非编码 RNA 分子,目前部分研究表明 lncRNA 对肺癌的发展过程具有一定影响. Tang 等^[10]发现一种新的没有被注释的 onco-lncRNA, onco-lncRNA-230,与正常组织比较,该基因在头颈部鳞状细胞癌 (HNSC)、肺腺癌 (LUAD)、肺鳞状细胞癌 (LUSC) 显著上调. Wu 等^[11]通过收集 TCGA-LUSC 的 lncRNA 与 mRNA 表达谱,建立 lncRNA-mRNA 共表达网路,最终识别出 lncRNA 相关的蛋白质相互作用模块,有助于理解 lncRNA 在肺癌中的调控作用. 以上方法说明了 lncRNA 对肺癌的发生发展有一定的影响,有助于研究肺癌的致病机理.

本文通过分析不同层面数据的集成模式,通过实验分析与验证,科学引入 lncRNA 表达谱数据,进一步扩大多组学数据集成方法的广度,挖掘候选基因构成的致癌基因模块,并对得到的结果从多个角度进行分析与验证,最终得到的 15 个基因能够良好地区分肿瘤样本与普通样本,并且具有良好的预后性能,可以作为预后标志物.

2 多组学数据

本文通过 TCGA 数据库收集与 LUSC 相关的多组学数据,包括 DNA 拷贝数变异、DNA 甲基化、基因表达,以及 miRNA 表达数据. 我们从 TCGA 数据库直接导出了包括 338 个样本的体细胞变异数据 (level 2). 通过去除沉默突变,获取到包括 5982 个突变基因集合. 通过 GISTIC 软件 (版本 2.0) 处理拷贝数变异数据并得到 337 个样本的 CNV 矩阵数据. 下载 JHU-USC HumanMethylation450 平台检测的 367 个样本的 DNA 甲基化数据. 获取 337 个样本中共 20503 个基因的表达数据和 387 个样本中 1043 个 miRNA 的表达数据. 并从 TANRIC 数据库下载 220 个样本中 12728 个 lncRNA 表达数据. 最终,确定 66 个包括所有多组学数据的 LUSC 样本.

miRNA-target 数据从 StarBase 获得,为减小数据假阳性,选定如下筛选标准: (1) 至少有 1 个 CLIP-Seq 试验支持该 miRNA 靶位点; (2) miRNA 和靶基因的表达至少在 3 种癌症中是负相关的 (pearson correlation: $\gamma < 0$, p -value < 0.05); (3) 该靶向关系至少在 5 个预测软件中的 1 个被预测出来 (TargetScan, PicTar, PITA, miRanda/mirSVR 和 RNA22). 本文从 TRANSFAC 数据库获取实验验证的 TF-gene 调控关系,并使用来自 UCSC 基因组浏览器的保守 TF 绑定位点. 人类蛋白质-蛋白质交互数据从以下数据库获取: BioGRID, BIND, HPRD, IntAct, MINT, MIPS, PDZBase, DIP 和 Reactome. 删除了冗余数据以及目前没有被实验验证或在文献中预测的交互.

lncRNA 与基因作用关系通过如下方式获取. 我们知道 miRNA 对基因具有调控关系,miRNA 同

图 1 (网络版彩图) LUSC 关键基因模块挖掘方法流程图
Figure 1 (Color online) Progress of core module mining method

样对 lncRNA 具有调控关系. 假设被同一个 miRNA 调控的基因与 lncRNA 具有交互关系, 对于被同一个 miRNA 调控的基因与 lncRNA, 通过计算它们在 66 个样本中表达值的 pearson 相关系数 (pearson correlation: $\gamma > 0$, $p\text{-value} < 0.05$) 来获取 lncRNA 与基因间的关系.

3 整体方法

本文提出了一种系统方法通过多组学异常调控网络来挖掘癌症中存在的键关键基因模块. 该方法的具体流程描述如图 1 所示.

3.1 构造关键候选基因的基因变异谱

首先, 通过 DNA 突变数据筛选基因得到产生突变的基因, 构造基因突变谱. 然后应用 GISTIC 处理 DNA 拷贝数变异数据来确定癌症样本中每个基因的拷贝数变异离散化数值 (-2, -1, 0, 1, 2), 得到 DNA 拷贝数变异谱. 通过组合基因突变谱与拷贝数变异谱数据, 生成二进制基因变异矩阵 M , 矩阵中每行对应一个基因, 每列对应一个样本, 当第 i 行基因在第 j 列样本中既发生拷贝数变异又发生基因突变时, m_{ij} 赋值为 1, 否则为 0.

下一步, 根据两个标准来筛选关键候选基因: (1) 该基因在拷贝数变异谱中产生变异的样本数量大于总样本数量的 10%; (2) 该基因在变异样本与非变异中差异表达. 第 2 个筛选标准需要集成基因表

达谱数据, 在该基因的变异样本与非变异样本之间使用 T 检验, 标准为错误发现率小于 0.01. 通过以上两个标准我们得到了关键候选基因的变异信息.

3.2 挖掘关键候选基因的异常调控基因集合

针对每个关键候选基因, 需要分别对其构建异常调控网络. 针对每个异常调控网络, 需要具备关键候选基因的变异样本与非变异样本的调控网络. 首先, 需要根据关键候选基因的变异样本信息, 将癌症样本分为两组. 然后针对每组样本, 通过构建线性回归模型来表示基因表达的改变.

在包含 G 个样本的特定条件下, 给定一个基因 g , 有 T 个 TFs ($TF_1, TF_2, \dots, TF_T$) 和 M 个 miRNAs ($miRNA_1, miRNA_2, \dots, miRNA_M$) 以及 L 个 lncRNA ($lncRNA_1, lncRNA_2, \dots, lncRNA_L$) 与 g 绑定. 通过如下公式训练线性回归模型:

$$GE_g \approx \omega_{CN} CN_g + \omega_{DM} DM_g + \sum_t \omega_{TF_t} TF_t + \sum_m \omega_{miR_m} miRNA_m + \sum_l \omega_{lnc_l} lncRNA_l, \quad (1)$$

其中 GE_g 是 G 个样本中 g 的表达等级, CN_g 是 g 的拷贝数出现次数, DM_g 是 g 的甲基化值, TF_t 是调控 g 的第 t 个转录因子表达等级, $miRNA_m$ 是以 g 为目标的第 m 个 miRNA 表达等级, $lncRNA_l$ 是对 g 具有影响的第 l 个 lncRNA 表达等级. ω_{CN} , ω_{DM} , ω_{TF_t} , ω_{miR_m} , ω_{lnc_l} 分别代表 CN_g , DM_g , TF_t , $miRNA_m$, $lncRNA_l$ 的回归系数. 对于样本中的基因, 只有在肿瘤样本与正常样本中差异表达 (R 语言 SAM 包 $fold-change[>2, <1/2]$ $fdr < 0.05$) 才会被用来训练该模型. 本文采用两层神经网络来训练该模型, 并得到模型中各个变量的权重, 只有权重大于特定阈值的调控因子才会被确定用于构建网络. 通过确定每个差异表达基因的调控因子, 构建关键候选基因在变异样本与非变异样本中的调控网络. 进一步, 通过对两个网络进行“异或”操作, 得到关键候选基因的异常调控网络. “异或”操作定义如下: 网络中的每条边代表一个调控关系, 对于任意调控关系, 若同时在两个网络中出现则丢弃该关系, 若仅出现一次, 则保留.

为获取关键候选基因异常调控的基因集合, 将关键候选基因以及异常调控网络中的基因映射到蛋白质交互网络, 寻找蛋白质交互网络中关键候选基因两步范围内出现的异常调控网络基因, 将这些基因作为关键候选基因的异常调控基因集合.

3.3 挖掘关键基因模块

本文通过基于密度聚类方式对关键基因进行数据挖掘, 来获取关键基因模块. 本文中关键基因与关键基因间的距离定义如下:

$$Distance_{i,j} = 1 - \frac{|DRGS_i \cap DRGS_j|}{|DRGS_i \cup DRGS_j|}, \quad (2)$$

其中 $Distance_{i,j}$ 表示第 i 个关键基因与第 j 个关键基因间的距离, $DRGS_i$ 和 $DRGS_j$ 分别表示第 i 个基因与第 j 个基因所影响的异常调控基因集合. 聚类算法的具体描述如算法 1 所示.

对于算法中的扫描半径以及邻域密度阈值两个参数, 本文采用 Yuan 等^[12] 提出的自动确定 DB-SCAN (density-based spatial clustering of application with noise) 算法参数的方法来确定, 得到两个参数值分别为 $Eps=0.726561679101$, $MinPts=4$, 在该参数下挖掘关键基因模块.

算法 1 基于密度的关键基因模块聚类算法
输入: Dataset: 一个包含 n 个对象的数据集
ϵ : 扫描半径参数
MinPts: 邻域密度阈值
输出: 具有簇标签的基因集合
方法:
1. 创建空间为 $n \times n$ 的二维矩阵 dis 用于存储关键基因间距离;
2. 计算 dataset 中关键基因间距离并保存到 dis 矩阵;
3. 对于 dataset 中的对象, 根据 dis 标记满足 ϵ 范围内密度大于 Minpts 的对象为 core point;
4. 标记 core point 在 ϵ 范围内的非 core point 对象为 border point;
5. 标记 dataset 中既不是 core point 也不是 border point 的对象为 noise point;
6. 标记 dataset 中所有对象为 unvisited;
7. for core point 中的每个对象 p //深度优先连接所有 core point
8. 创建栈 Stack;
9. if p 是 visited
10. Continue;
11. 将 p 标记为 visited 压入 Stack
12. while(Stack 不为空)
13. $v \leftarrow$ Stack 弹出栈顶;
14. for v 邻域内的每个对象 q
15. if q 是 visited
16. Continue;
17. 将 q 纳入 p 所在 cluster;
18. 将 q 标记为 visited;
19. 将 q 压入 Stack;
20. end for
21. end while
22. end for
23. for border point 中的每个对象 b
24. 将 b 纳入 ϵ 范围内任意 core point 所属 cluster;
25. end for
26. 输出 dataset 中 core point 与 border point 对象以及其对应的 cluster.

4 结果

4.1 lncRNA 有助于提升模型精度

Li 等^[7] 证明通过整合多组学数据对基因表达数据进行线性回归拟合的有效性, 但并未考虑 lncRNA. 为验证增加 lncRNA 有助于提升本文模型的预测精度, 本文从 66 个样本中随机抽取 50 个样本作为训练集, 剩余样本作为验证集, 对于加入与未加入 lncRNA 两种情况分别进行 100 次交叉验证, 取验证集的均值与原始样本均值进行对比, 计算 Pearson 相关系数. 此外, 还进行了 100 次随机数据与实际数据对比. 实验结果显示, 加入 lncRNA 有助于提升模型预测的准确率. 具体实验结果如图 2

图 2 (网络版彩图) lncRNA 实验结果对照箱式图

Figure 2 (Color online) lncRNA experimental result. (a) Results while including lncRNA and random data; (b) results while excluding lncRNA data

所示.

4.2 关键基因模块

通过本文提出的方法, 我们得到了由 15 个基因组成的关键基因模块, 包括 DNAJB11, SMC4, DLG1, PRKCI, TNK2, DROSHA, ECT2, CCDC50, SKIL, TP53, DAB2, AP2M1, DVL3, CDKN2A, AHSB. 其中 DAB2, DROSHA, CDKN2A, TP53 这 4 个基因与 LUSC 相关 (从 Cosmic, OMIM, HuGE 和 GAD 中下载 LUSC 基因列表). 这些基因以及与其相关的转录因子, miRNA 以及 lncRNA 如图 3 所示. 该图中的红色圆形实体代表关键基因, 蓝色圆形实体代表转录因子, 绿色三角形代表 miRNA, 黄色正方形代表 lncRNA. 关键基因模块内部各个基因间的关系以及对关键基因具有调控作用的转录因子、miRNA 以及 lncRNA 的个数如图 4 所示. 该图中外圈实体代表 15 个关键基因, 其中红色、蓝色、黄色、绿色分别代表该基因属于 3, 5, 9, 17 号染色体. 内圈柱状图代表该基因调控因子个数, 土黄色、浅蓝色、亮紫色分别代表转录因子、miRNA、lncRNA. 图中紫色连线代表基因与基因间关联关系. 各关键基因染色体信息如表 1 所示.

4.3 区分正常与肿瘤样本

为验证关键基因模块在 LUSC 发展过程中的关键作用. 对包含 337 个 TCGA-LUSC 样本的基因表达谱数据进行层次聚类, 距离度量采用欧式距离, 计算类与类之间的平均距离. 聚类结果如图 5 所示. 该图中的第一行, 红色代表该样本为正常样本, 绿色代表肿瘤样本. 剩余部分每行代表一个基因, 每列代表一个样本, 表示特定基因在特定样本下的表达值, 由红到蓝表示表达值的高低. 根据图中所示, 发现这 15 个基因能够良好区分癌症样本与正常样本.

4.4 功能与通路

为进一步确认 15 个关键基因的作用, 应用 DAVID 软件进行 Gene Ontology 功能和 KEGG 通路

图 4 关键基因模块

Figure 4 Core gene module

表 1 关键基因染色体信息

Table 1 Chromosome information of core genes

Gene	Chromosome	Start point	End point
TP53	hs17	7668402	7687550
CDKN2A	hs9	21967752	21995043
DROSHA	hs5	31400494	31532175
DAB2	hs5	39371674	39425233
SMC4	hs3	160399304	160434962
PRKCI	hs3	170222432	170305982
SKIL	hs3	170357678	170396849
ECT2	hs3	172750682	172829273
DVL3	hs3	184155311	184173614
AP2M1	hs3	184174846	184184091
DNAJB1	hs3	186570676	186585800
AHSG	hs3	186612928	186621318
CCDC50	hs3	191329082	191398670
TNK2	hs3	195863364	195909009
DLG1	hs3	197042560	197299272

图 5 根据关键基因区分正常与肿瘤样本实验结果

Figure 5 Result of distinguish normal and tumor samples according to core genes

图 6 功能与通路分析结果

Figure 6 Result of function and pathway analysis

一部分, 虽然其不在已知的生物通路中, 但去掉它们会在一定程度上影响生存分析结果.

我们还在 GEO 数据库中下载了 GSE8894, GSE17710 这两组数据, 并分别进行生存分析, 同样发现加入这 3 个基因比不加入的效果更好, 分析结果如图 8 所示. 因此证明, 在 TCGA 数据上的分析结果并非偶然, 这 3 个基因确实在 LUSC 生存分析中具有重要作用.

5 讨论

本文提出一种基于多组学生物网络的癌症关键模块挖掘方法, 通过集成多组学生物数据, 构造关键基因的异常调控网络, 挖掘与 LUSC 相关的关键基因模块以及它们所影响的异常调控基因集合. 最终得到由 15 个基因组成的关键基因模块. 其中 DAB2, DROSHA, CDKN2A, TP53 与 LUSC 相关 (属

表 2 功能 ID 与功能名称
Table 2 Function ID and function name

ID	Term
GO:0035556	Intracellular signal transduction
hsa04144	Endocytosis
GO:0032147	Activation of protein kinase activity
hsa05203	Viral carcinogenesis
GO:0043065	Positive regulation of apoptotic process
GO:0046677	Response to antibiotic
GO:0071479	Cellular response to ionizing radiation
GO:0090004	Positive regulation of establishment of protein localization to plasma membrane
GO:0071158	Positive regulation of cell cycle arrest
GO:0042326	Negative regulation of phosphorylation
hsa04390	Hippo signaling pathway
GO:0045197	Establishment or maintenance of epithelial cell apical/basal polarity
GO:0090399	Replicative senescence
GO:0045893	Positive regulation of transcription, DNA-templated
GO:0007050	Cell cycle arrest
GO:1903077	Negative regulation of protein localization to plasma membrane
hsa05166	HTLV-I infection
GO:0070830	Bicellular tight junction assembly

于 Cosmic, OMIM, HuGE 和 GAD 中 LUSC 基因并集). DNAJB11 被证明在鳞状细胞癌的抗药性中发挥着重要作用^[13]. Massion 等^[14] 为寻找吸烟导致癌症的生物标签, 通过进行 GO 与 KEGG 通路分析, 将 SMC4 作为与吸烟相关的标签, 该基因在病人样本中产生显著的拷贝数变异, 参与 DNA 复制、染色体分离等. DLG1 被证明为乳腺导管癌 (mammary ductal carcinoma) 的抑癌基因^[15]. PRKCI 被证明为 LUSC 致癌基因^[16]. TNK2 基因扩增是一种新的针对胃癌患者的不良预后预测因素^[17]. ECT2 是人类癌症中的致癌基因^[18]. 基因敲除研究显示 CCDC50 是套细胞淋巴瘤与慢性淋巴细胞白血病的候选基因^[19]. SKIL 是非小细胞肺癌与乳腺癌的癌症驱动基因^[20]. 经过去甲基化治疗, AP2M1 在 4 个肝癌细胞线 (Huh6, PLC/PRF/5, HLE, Huh7) 中明显的下调^[21]. DVL3 是非小细胞肺癌 Wnt 通路的激活者^[22]. AHSB 是与肺癌相关的基因^[23]. 这 15 个基因中有 10 个与肺癌相关, 其中 9 个与非小细胞肺癌相关, 7 个与 LUSC 相关.

在 NCBI 调查本文中的关键基因, 发现 15 个关键基因中有 11 个位于 3 号染色体. 人类 3 号染色体携有高密度与癌症相关的基因, 自 20 世纪 90 年代中期以来成为生命科学界的研究重点^[24]. 本文的发现在一定程度上说明了人类 3 号染色体与癌症的相关性. 为癌症与染色体层面上的研究提供了材料基础.

与以往方法相比较, 本文首次引入 lncRNA, 并发现引入 lncRNA 可以提升线性回归模型的预测能力, 在前人的基础上进一步扩大了多组学数据集成方法的广度. 本文还提出一种新的基于密度的关键基因模块聚类算法, 该聚类方法成功将关键基因这种抽象数据映射到聚类空间中, 使得关键基因之间的距离能够被度量.

图 7 (网络版彩图) TCGA 生存分析结果

Figure 7 (Color online) TCGA survival analysis result. (a) 15-genes; (b) 12-genes

图 8 (网络版彩图) GEO 生存分析结果

Figure 8 (Color online) GEO survival analysis result. (a) 15-genes GSE8894; (b) 15-genes GSE17710; (c) 12-genes GSE8894; (d) 12-genes GSE17710

6 结束语

本文利用数据挖掘技术分析多源异构生物数据, 挖掘基因与癌症的关系. 通过集成多组学数据构建生物网络模型, 综合利用了多组学数据中富含的信息, 基于此进一步挖掘关键基因模块以及基因通路, 能够发现单组学数据难以获得的深层次联系. 这些关系有助于为癌症研究提供新的视角, 具有积极的现实意义.

目前, 大部分基于数据集来挖掘癌症中潜藏信息的方法都是基于静态网络的, 没有考虑到时间

维度. 但生物网络模块的功能是随时间动态演变的, 随着外部条件的改变, 模块的功能也会随之改变, 涉及到基因与基因间交互的动态性. 接下来我们将研究动态生物网络中的功能模块探测及其演变模式分析方法, 期望能够揭示生命代谢过程的结构以及演变的事件、模式和规则.

参考文献

- 1 Reuter J A, Spacek D V, Snyder M P. High-throughput sequencing technologies. *Mol Cell*, 2015, 58: 586–597
- 2 Chuang H Y, Lee E, Liu Y T, et al. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 2007, 3: 141–150
- 3 Zhang S H, Liu C C, Li W Y, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*, 2012, 40: 9379–9391
- 4 Hoadley K A, Yau C, Wolf D M, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 2014, 158: 929–944
- 5 Yang Y, Han L, Yuan Y, et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun*, 2014, 5: 3231
- 6 Han L, Yuan Y, Zheng S Y, et al. The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun*, 2014, 5: 3963
- 7 Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*, 2014, 10: e1003908
- 8 Ping Y, Deng Y, Wang L, et al. Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multi-dimensional genomic data. *Nucleic Acids Res*, 2015, 43: 1997–2007
- 9 Wang Y F, Liu L L, Jin H F, et al. Study on expression profile of long non-coding rna in gastric cancer cell lines under hypoxia. *J Mod Oncol*, 2013, 21: 225–228
- 10 Tang C Y, Silafisher J M, Dang H X, et al. Abstract 971: a novel long noncoding rna, onco-lncrna 230, induces apoptosis and invasion in lung squamous cell carcinoma. *Cancer Res*, 2016, 76: 971
- 11 Wu C H, Hsu C L, Lu P C, et al. Identification of lncrna functions in lung cancer based on associated protein-protein interaction modules. *Sci Rep*, 2016, 6: 35939
- 12 Yuan F, Meng Z H, Yu G. An improved dbSCAN clustering algorithm. *J Comput Res Dev*, 2005, 42: 50–54 [袁方, 孟增辉, 于戈. 对 dbSCAN 聚类算法的改进. *计算机研究与发展*, 2005, 42: 50–54]
- 13 Wen J, Zheng B, Hu Y, et al. Comparative proteomic analysis of the esophageal squamous carcinoma cell line ec109 and its multi-drug resistant subline ec109/cddp. *Int J Oncol*, 2010, 36: 265–274
- 14 Massion P P, Zou Y, Chen H, et al. Smoking-related genomic signatures in non-small cell lung cancer. *Am J Respir Crit Care Medicine*, 2008, 178: 1164–1172
- 15 Fuja T J, Lin F, Osann K E, et al. Somatic mutations and altered expression of the candidate tumor suppressors *csnk1 epsilon*, *dlg1*, and *edd/hhyd* in mammary ductal carcinoma. *Cancer Res*, 2004, 64: 942–951
- 16 Justilien V, Walsh M P, Ali S A, et al. The *prkci* and *sox2* oncogenes are coamplified and cooperate to activate hedgehog signaling in lung squamous cell carcinoma. *Cancer Cell*, 2014, 25: 139–151
- 17 Shinmura K, Kiyose S, Nagura K, et al. *Tnk2* gene amplification is a novel predictor of a poor prognosis in patients with gastric cancer. *J Surg Oncol*, 2014, 109: 189–197
- 18 Fields A P, Justilien V. The guanine nucleotide exchange factor (*gef*) *ect2* is an oncogene in human cancer. *Adv Enzyme Regul*, 2010, 50: 190–200
- 19 Farfing A, Engel F, Seiffert M, et al. Gene knockdown studies revealed *ccdc50* as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia. *Leukemia*, 2009, 23: 2018–2026
- 20 Hagerstrand D, Tong A, Schumacher S E, et al. Systematic interrogation of 3q26 identifies *tloc1* and *skil* as cancer drivers. *Cancer Discov*, 2013, 3: 1044–1057
- 21 Arai M, Yokosuka O, Hirasawa Y, et al. Sequential gene expression changes in cancer cell lines after treatment with the demethylation agent 5-aza-2'-deoxycytidine. *Cancer*, 2006, 106: 2514–2525
- 22 Uematsu K, He B, You L, et al. Activation of the wnt pathway in non small cell lung cancer: evidence of dishevelled overexpression. *Oncogene*, 2003, 22: 7218–7221

- 23 Dowling P, Clarke C, Hennessy K, et al. Analysis of acute-phase proteins, ahsg, c3, cli, hp and saa, reveals distinctive expression patterns associated with breast, colorectal and lung cancer. *Int J Cancer*, 2012, 131: 911–923
- 24 Muzny D M, Scherer S E, Kaul R, et al. The DNA sequence, annotation and analysis of human chromosome 3. *Nature*, 2006, 440: 1194–1198

A method for mining core modules of cancer based on multi-omics biological network

Maozu GUO¹, Xuejian WU², Ning ZHAO³, Xiaoyan LIU² & Chunyu WANG^{2*}

1. *School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;*

2. *School of Computer Science and Technology, Harbin institute of technology, Harbin 150000, China;*

3. *School of Life Science and Technology, Harbin institute of technology, Harbin 150000, China*

* Corresponding author. E-mail: chunyu@hit.edu.cn

Abstract Cancer is closely linked to factors such as human living environment, individual genetic factors, etc. Because of the serious threat that cancer brings to human health, numerous scientific institutions around the world are engaged in cancer research to understand its pathogenesis. With the advent of next-generation sequencing technology, it will be more convenient to find important information about cancer and the relationships in genome. This paper considers multi-omics data from a data integration perspective. It takes lncRNA omic data into consideration and constructs a biological network model to mine cancer-associated core gene modules through a cluster method. We present a systematic approach to the identification core gene modules that can lead to the occurrence of cancer. We apply this approach to lung squamous cell carcinoma and find core gene modules containing 15 genes that have strong relationship with cancer by analyzing their functions and pathways. We also distinguish high-risk and low-risk groups by survival analysis. The results show that our approach can identify core gene modules and their dysregulated genes by integrating multi-omics biological data, which is useful in cancer research.

Keywords multi-omics data, biological network model, DBSCAN, lncRNA, dysregulation

Maozu GUO was born in 1966. He received his B.S. degree from Harbin Engineering University, China, in 1988 and Ph.D. degree in computer science and technology from Harbin Institute of Technology. He is a professor of Civil Engineering and Architecture, Beijing University. His research interests include bioinformatics, machine learning, and data mining.

Xuejian WU was born in 1993. He received his B.S. degree from Harbin Engineering University, China, in 2015. He is currently working toward a master's degree at Harbin Institute of Technology. His research interests include bioinformatics and machine learning.

Ning ZHAO was born in 1989. She received her B.S. and M.S. degrees in bioinformatics from Harbin Medical University, China, in 2013 and 2016, respectively. She is currently working toward a doctorate degree at Harbin Institute of Technology. Her current research interests include cancer genetics and bioinformatics.

Chunyu WANG was born in 1979. He received his B.S., M.S., and Ph.D. degrees in computer science and technology from Harbin Institute of Technology in 2002, 2005, and 2015, respectively. He is an associate professor at the Department of Computer Science and Technology. His interests include bioinformatics and machine learning.