Computer Science and Application 计算机科学与应用, 2019, 9(9), 1792-1796
Published Online September 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.99200

Hans汉斯

Breast Cancer Drugs Repositioning Based on Mutual Information

Yifan Hao, Guang Yang

Mathematics and Systems Science, Shenyang Normal University, Shenyang Liaoning
Email: 897120417@qq.com

Received: Sep. 3rd, 2019; accepted: Sep. 18th, 2019; published: Sep. 25th, 2019

Abstract

In order to solve the problem of drug relocation for breast cancer, first, breast cancer gene expression data are gotten in the TCGA database, and then through the mutual information algorithm, the feature genes are extracted, finally, through connectivity map analysis results, the compared drugs are ranked in descending order according to the negative correlation scores, getting gliquidone and imatinib drugs that may have treatment effects on breast cancer. Compared with traditional drugs, targeted therapies that have therapeutic effects on cancer genes are more targeted by extracting characteristic genes. Screening anti-breast cancer drugs through retargeting method not only greatly reduces the cycle of new drug development, but also reduces the economic cost. Extracting feature genes based on mutual information algorithm provides a new way for drug relocalization.

Keywords

Drug Repositioning, Breast Cancer, Mutual Information, Connectivity Map

基于互信息算法的抗乳腺癌药物重定位分析

郝逸凡，杨　光

沈阳师范大学数学与系统科学学院，辽宁 沈阳
Email: 897120417@qq.com

收稿日期：2019年9月3日；录用日期：2019年9月18日；发布日期：2019年9月25日

摘　要

针对抗乳腺癌药物重定位问题，首先在TCGA数据库中获取乳腺癌基因表达数据，然后通过互信息算法提

文章引用: 郝逸凡, 杨光. 基于互信息算法的抗乳腺癌药物重定位分析[J]. 计算机科学与应用, 2019, 9(9): 1792-1796.
DOI: 10.12677/csa.2019.99200

取特征基因，最后通过connectivity map分析结果，将比对出的药物按照负相关的分值降序排列，得到gliquidone (格列喹酮)，Imatinib (格列卫)等可能对乳腺癌有治疗效果的药物。与传统药物相比，通过提取特征基因，找出对癌症基因有治疗效果的靶向治疗更具有针对性，通过重定位的方法筛选出抗乳腺癌药物不仅大大减少了新药开发的周期，还降低了经济成本。基于互信息算法提取特征基因为药物重定位提供了新的途径。

关键词

药物重定位，乳腺癌，互信息，Connectivity Map

1. 引言

乳腺癌是发生在乳腺腺上皮组织的恶性肿瘤，已经成为威胁女性身心健康的常见肿瘤。全球乳腺癌的发病率目前一直呈上升趋势。中国虽不是乳腺癌的高发国家，但近年我国的乳腺癌发病率的增长速度却高出高发国家 1~2 个百分点，情况仍不容乐观。随着对乳腺癌生物学行为认识的不断深入，以及治疗理念的转变与更新，乳腺癌的治疗进入了综合治疗时代，形成了乳腺癌局部治疗与全身治疗并重的治疗模式。乳腺癌的外科手术包括乳腺和腋窝淋巴结两部分。乳腺手术有保留乳房手术(保乳手术)和全乳房切除术。保乳手术有严格的手术适应症，目前还做不到所有的乳腺癌患者都能进行保乳手术。对不适合保乳手术的乳腺癌患者还需要切除乳房。

鉴于国内现有的医疗水平，医生会根据肿瘤的分期和患者的身体状况，酌情采用手术、放疗、化疗、内分泌治疗、生物靶向治疗及中医药辅助治疗等多种手段。目前，分子靶向治疗已成为肿瘤治疗的研究热点，为乳腺癌的治疗也提供了新的思路和方向。利用基因表达谱等组学技术发现抗乳腺癌的药物靶标可作为一个重要手段。但新药开发是一个耗时费力的高风险过程，充分发掘已有药物的新用途，对药物进行重定位，备受生物医药产业和学者们的青睐[1] [2] [3]。

药物重定位又称老药新用，指对曾经用于临床的药物新适应症的发现、确认和应用。包括对处于临床研究阶段或已批准上市的药物进行重定位、重定用途、重评价和重新定位治疗方向等[4]。推动一个新药物上市通常需要 13~15 年，其成本平均需要 20~30 亿美元，且处于上升趋势。如果对已有药物进行研究，一旦它们拥有不同的医疗用途，这将是一个巨大的未开发资源。“药物重定位”可以跳过临床 I 期，相比于新药物大大地缩减研究成本和投入时间。到目前为止，从已知的药物中发现新的适应症，成功重定位的药物已经有 100 多种。如何从已知药物中发现对于乳腺癌有治疗效果的药物是本文探讨的问题。

Connectivity map 是利用小分子药物、基因表达与疾病相互关联的生物应用数据库。通过基因表达谱建立基因、疾病和药物三者的关联性，并快速利用基因表达谱的数据比对出与疾病高关联性的药物。近年来的研究趋势表明：将 cmap 基因表达谱数据库应用于疾病治疗与药物开发领域，可提供越来越精确的方向。在药物开发方面，利用基因表达谱的数据在 cmap 数据库中快速比对出与疾病高关联性的药物。目前已经有学者成功地利用 cmap 验证了抗溃疡药可以用于治疗肺癌，抗癫痫药物可以用来治疗炎症性肠道疾病，抗哮喘药物可以用来预防白内障等。如何将这种方法应用在抗乳腺癌药物的领域里是本文研究的问题。

本文首先从 TCGA 数据库中获取乳腺癌与癌旁的基因表达数据，利用 R 软件将数据进行预处理；然后利用互信息算法将与乳腺肿瘤密切相关的特征基因筛选出来；最后通过 connectivity map 数据库分析，检索出具有与肿瘤基因相反的基因标签的药物。Gliquidone (格列喹酮)作为一种降血糖药，用于非胰岛素依赖型糖尿病的治疗，经分析比对得到的负相关分值最高，表明对于乳腺癌可能具有较好的治疗效果。Imatinib (格列卫)、levomepromazine (左米丙嗪)、erastin 等化合物也具有较高的负相关分值，表明极可能对乳腺癌有治疗效果。

2. 数据和方法

2.1. 基因表达数据

TCGA 是美国国家癌症研究所(National Cancer Institute)和美国人类基因组研究所(National Human Genome Research Institute)共同监督的一个项目，旨在应用高通量的基因组分析技术，以帮助人们对癌症有个更好的认知，从而提高对于癌症的预防、诊断和治疗能力。作为目前最大的癌症基因信息数据库，TCGA 数据库主要收录各种人类癌症(包括亚型在内的肿瘤)的临床数据、基因组变异和 mRNA 表达等数据，是癌症研究者十分重要的数据来源[5]。本文的乳腺癌基因表达数据来自 TCGA 数据库，共获得乳腺癌与癌旁的基因表达数据，其中包括 534 个患病样本和 63 个健康样本，包括 17815 条基因(https://cancergenome.nih.gov/)。

2.2. 特征基因提取

对于复杂的基因关系，熵和互信息的方法能有效抓住基因与基因之间的关联性，能有效的提取出复杂疾病的致病基因[6]。熵是对不确定性的度量，在信息论中，熵是用来衡量一个随机变量出现的期望值。设基因变量 $X=\left[x_1,x_2,\cdots,x_n\right]$ 是一个基因表达模式[7]，基因变量 X 的熵表示该模式所包含的信息量公式为：

$$H\left(X\right)=-\sum_{x\in X}p\left(x\right)\log_2^{p(x)}$$

互信息是信息论中的一种有用的信息度量。可以看成是一个随机变量中包含的关于另一个随机变量的信息量。对于两个随机变量 X 和 Y，其互信息公式为：

$$I\left(X,Y\right)=\sum_{x\in X}\sum_{y\in Y}p\left(x,y\right)\log_2\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}$$

传统的特征基因提取方法通常只注重单个基因的表达，而忽略了基因之间的关联性。针对这种情况可以采取基于互信息算法的特征基因提取，如果互信息值比较大，说明这两个基因之间的关联性比较大，即在生物学上的联系比较紧密。计算出每条基因的信息熵，信息熵越大，证明该条基因在样本中拥有较大的信息量，对样本的影响也就越大。所以基于信息熵的角度，将每条基因的信息熵降序排列，取前 5000 个基因，计算其在患病样本和健康样本中的互信息值，得到两个互信息矩阵矩阵，即 I_c 和 I_n。

在健康样本中和其他基因关联较小即互信息值较小，在患病样本中该基因又与其他基因具有较大的关联性即互信息值较大的基因为从失联到关联状态下的基因，则可认为此类基因为特征基因。提取特征基因的关键在于找出合适的阈值 T_c 和 T_n 使得特征基因数目不会过多，也不会太少，经过计算从失联到关联状态的特征基因的理想阈值为 $T_c=0.66$ 和 $T_n=0.62$，从而得到实对称矩阵，对其按行求和并将和值降序排列，和值越大证明该基因在样本中与越多的基因相关联，和值为 0 则代表该基因并不与其他基因有

关联[8]。根据上述步骤获得从失联到关联状态下的特征基因 656 条。

2.3. cmap 分析

Connectivity map 是一个基因表达谱数据库，其利用小分子药物，基因表现与疾病相互关联的生物应用数据库。每一种药物分子会以不同浓度处理在不同的细胞株并处理不同的时间点。基因表达谱数据区分成正向调控基因群和负向调控基因群进行分析，以运算基因图谱的相似程度为主，最后给予分数。分数越接近 1 代表两者的药物分子为正相关，反之，与负向调控基因群的基因图谱相近之药物分子，分数呈为负值，代表两者的药物分子为负相关。以基因表达谱为所建立之基因，疾病与药物的关联性。可以快速利用基因表达谱的数据比对出与疾病高关联性的药物。近年来的研究趋势也显示出利用 cmap 基因表达谱数据库应用在疾病治疗与药物开发的领域上，可提供越来越精确的方向。目前 cmap 第二版已经发展成收录了 1309 种药物表达谱的成熟体系，理论上讲，与疾病和药物相关的任何基因表达数据都可以在 cmap 数据库中进行高效率的查询比对，从数据库揭示药物、基因和疾病三者之间潜在的联系[9]。

通过 R 软件将筛选出的特征基因分为上调基因 294 个和下调基因 362 个。将上调基因和下调基因作为检索标签，存为.grp 文件，检索 cmap 数据库[10]。将乳腺基因表达标签与药物处理基因标签进行统计比较。依据表达谱的相似性给每个乳腺癌-药物配对计算一个分值，如果分值为负数，则表明这种药物与癌症基因有相反的基因标签，即可能对乳腺癌具有较好的治疗效果。所以在检索的过程中，删除试验次数较少的药物($n < 4$)，关注药物得分 Mean 分值为负值的药物。

3. 结果分析

通过 connectivity map 的分析结果如表 1 所示可以看出负相关分值最高的是 gliquidone (格列喹酮)，分值为−0.618，它是一种降血糖药，用于非胰岛素依赖型糖尿病的治疗，表明对于乳腺癌可能具有较好的治疗效果；表中还可以看出排在后面的 Imatinib (格列卫)、levomepromazine (左米丙嗪)、erastin 等化合物也具有较高的负相关分值，表明极可能对乳腺癌有治疗效果。其中表中的 tamoxifen (他莫昔芬)，用于治疗晚期乳腺癌和卵巢癌。排在它上面的药物最后的药物检索分值的负相关性均高于它，所以这几种药物很可能与治疗乳腺癌有关。

Table 1. Candidate anti-breast cancer drugs screened by the connectivity map database
表 1. Connectivity map 数据库筛选出的候选抗乳腺癌药物

cmap name	Mean	*n*	Enrichment
gliquidone	−0.703	5	−0.711
Imatinib	−0.69	4	−0.624
levomepromazine	−0.572	4	−0.536
erastin	−0.448	3	−0.509
tamoxifen	−0.07	7	−0.229

注：Mean 表示药物检索得分值，*n* 为药物在 cmap 数据库中重复试验的次数，enrichment 为乳腺癌症基因标签与药物基因标签相似的聚合度。

4. 结论

本文通过互信息算法提取乳腺癌中的特征基因，利用 connectivity map 数据库将基因与药物进行比对打分，最后得到与治疗乳腺癌有关的药物。与传统药物相比，利用互信息算法提取出含有较多信息的致病基因，通过 cmap 数据库筛选出针对这些特征基因有疗效的靶向治疗，不仅能节约新药开发的时间和经

济成本，还拥有比传统治疗更好的治疗效果。但具体药效有待临床进一步验证。基于 R 语言中的互信息算法提取到癌症特征基因为抗癌药物重定位提供了新的途径。

基金项目

国家自然科学基金资助项目(61703290)，辽宁省科技厅自然科学基金资助项目(20180550133)；辽宁省教育厅科学技，术研究项目(LQN201710)。

参考文献

[1] Fu, C.H., *et al.* (2013) DrugMap Central: An On-Line Query and Visualization Tool to Facilitate Drug Repositioning Studies. *Bioinformatics*, **29**, 1834-1836. https://doi.org/10.1093/bioinformatics/btt279

[2] Zhao, K. and So, H.C. (2018) Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE Journal of Biomedical & Health Informatics*, **23**, 1304-1315. https://doi.org/10.1109/JBHI.2018.2856535

[3] 汪浩，王海平，吴信东，刘琦．药物-疾病关系预测：一种推荐系统模型[J]．中国药理学通报，2015，31(12): 1770-1774.

[4] 李苗苗．基于 XG-B00ST 和多数据源的药物重定位预测[J]．软件导刊, 1-5[2019-09-17].

[5] Kanehisa, M., Goto, S., Kawashima, S., *et al.* (2004) The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research*, **32**, D277.

[6] 孙啸，陆祖宏，谢建明．生物信息学基础[M]．北京：清华大学出版社, 2005: 283-315.

[7] 孔薇，支星，牟晓阳．基于互信息的显著基因提取及转录调控网络构建[J]．计算机应用与软件，2016，33(6): 235-239.

[8] 仲如星，孔薇．基于互信息和距离相关性算法的乳腺癌信号转导通路串扰[J]．科学技术与工程，2017，17(29): 205-211.

[9] 张晓芳，康永波，苏君鸿，孔祥阳．Connectivity map 技术在中药研究中的应用[J]．浙江大学学报(农业与生命科学版), 2016, 42(5): 543-550.

[10] 任磊，郑冰清，曲新艳，杨云，王建平，颜耀东．基因表达标签相似性比较的抗辐射药物重定位研究[J]．临床药物治疗杂志, 2017, 15(3): 19-24.