Advances in Clinical Medicine 临床医学进展, 2019, 9(9), 1098-1104
Published Online September 2019 in Hans. http://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2019.99169

Hans汉斯

Prostate Cancer Drugs Repositioning Based on Entropy and Fold Change Algorithm

Yifan Hao, Guang Yang

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

Received: Sep. 2nd, 2019; accepted: Sep. 16th, 2019; published: Sep. 23rd, 2019

Abstract

Drug repositioning research not only greatly reduces the cycle of new drug development, but also reduces the economic cost. Prostate cancer is a cancer with high mortality in the world. In recent years, the research on single gene is not enough to analyze the complex pathogenesis of cancer. In this paper, information entropy and Fold Change (FC) algorithm in R software were applied to analyze the gene expression data of prostate cancer and extract the characteristic genes of prostate cancer with a large number of pathogenic gene information. Then several drugs for prostate cancer were identified by cmap analysis. The gene expression data used in this paper were from TCGA, and 2788 up-regulated genes and 2222 down-regulated genes were obtained through FC algorithm. And characteristics of these genes screened 200 select genes with significant information as retrieval label cmap (https://portals.broadinstitute.org/cmap/) analysis; finally, the negative correlation score of cmap was ranked in descending order, and several possible effective drugs for the treatment of prostate cancer were analyzed and determined. Compared with the traditional method of extracting characteristic genes, the combination of information entropy and Fold Change algorithm can reduce the computational steps and amount of computation, with low cost, low time consumption and high accuracy. This method is effective and feasible.

Keywords

Drug Repositioning, Prostate Cancer, Entropy, Fold Change, Connectivity Map

基于熵和Fold Change算法的抗前列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2日；录用日期：2019年9月16日；发布日期：2019年9月23日

文章引用: 郝逸凡, 杨光. 基于熵和 Fold Change 算法的抗前列腺癌药物重定位分析[J]. 临床医学进展, 2019, 9(9): 1098-1104. DOI: 10.12677/acm.2019.99169

摘　要

本文针对抗前列腺癌药物重定位问题，首先将R软件中的信息熵和差异表达倍数Fold Change算法结合起来，分析前列腺癌的基因表达数据，并提取含有大量致病基因信息的前列腺癌特征基因，通过FC算法得到2788条上调基因和2222条下调基因；然后对这些特征基因各筛选出200条具有显著信息的特征基因作为检索标签进行cmap (https://portals.broadinstitute.org/cmap/)分析；最后通过cmap负相关分数降序排列，分析确定几种可能对治疗前列腺癌有效的药物。与传统的提取特征基因方法相比，信息熵和Fold Change算法相结合能减少计算步骤和计算量，成本低、耗时少、准确度高。该方法有效可行。

关键词

药物重定位，前列腺癌，信息熵，差异表达倍数，Cmap

1. 引言

前列腺癌是发生在男性前列腺组织中的恶性肿瘤，是前列腺腺泡细胞异常无序生长的结果。前列腺癌发病率的高低与地理和种族的差异性有关。在欧美一些发达国家和地区，它是男性最常见的恶性肿瘤，死亡率排在各种癌症的第二位；在亚洲，虽然发病率低于其他西方国家，但是近几年也呈迅速上升趋势。临床上前期主要采用雄激素剥夺疗法(ADT)治疗前列腺癌，然而几乎所有患者最终都会发展为致命性的去势抵抗型前列腺癌(CRPC)。虽然 FDA (美国食品药品管理局)批准的第二代抗雄激素药物如 Enzalutamide (恩杂鲁胺)和 Abiraterone (阿比特龙)等对缓解疾病进展具有一定的功效，但患者很快就会出现临床耐药。因此，临床上迫切需要治疗前列腺癌的特效药。

鉴于国内现有的医疗水平，针对前列腺癌仅能通过常规手术治疗、内分泌及化学药物疗法来提高患者的生活质量，但提高患者的生存期依旧是一个难题。目前，分子靶向治疗已成为肿瘤治疗的研究热点，为前列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前列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前列腺癌有治疗效果的药物是本文探讨的问题。

信息熵是信息论中用于度量信息量的一个概念。对于基因来说，它的信息熵越大，所包含的信息量就越大，具有的生物学意义就越大，计算出基因的信息熵再求其互信息值来提取特征基因是常用的特征基因提取方法。但是用互信息方法提取上调基因和下调基因的过程中，对阈值的选取要求比较高，阈值选取不当，会导致上调基因和下调基因选取不当。

Fold Change (简称 FC)是差异表达倍数，对于基因表达数据来说，计算其 logFC 值来筛选特征基因，如果 logFC 值为正则为上调基因；如果为负则为下调基因。若只用 FC 值来提取特征基因，计算所有基因的 FC 值再取对数则计算量大且时间成本高。但如果将两个方法结合，先利用信息熵提取出一些致病信息量大的特征基因，再通过计算它们的 FC 值再筛选出上调基因和下调基因，不仅省去选取阈值的步骤，而且只需计算信息熵大的基因的 FC 值，大大缩减计算步骤和计算量，耗时低，结果也更精确。如何在计算信息熵的基础上结合 FC 值提取出前列腺癌的特征表达基因是本文探讨的问题。本文首先从 TCGA 数据库中获取前列腺癌与癌旁的基因表达数据，利用 R 软件将数据进行标准化预处理；然后利用互信息算法算出每条基因的信息熵，从大到小排列，将与前列腺肿瘤密切相关即信息熵值最大的前 5000 个特征基因筛选出来，再利用计算其 logFC 值，筛选出上调基因和下调基因；最后通过 cmap 数据库[5]分析，检索出具有与肿瘤基因相反的基因标签的药物。gliclazide (格列齐特)作为一种治疗中型糖尿病的药物，与胰岛素合用治疗胰岛素依赖型糖尿病，可减少胰岛素用量，经分析比对得到的负相关分值最高，表明对于前列腺癌可能具有较好的治疗效果。Luteolin (木犀草素)、phenoxybenzamine (竹林胺)、naloxone (盐酸纳洛酮)等化合物也具有较高的负相关分值，表明极可能对前列腺癌有治疗效果。

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 数据库的全面不止体现在众多的癌型上，还体现在多组学数据，主要收录基因表达数据、mRNA 表达数据、拷贝数变异、DNA 甲基化、SNP 等等，是癌症研究者十分重要的数据来源。TCGA 数据库最大的优势就是包含各种人类癌症(包括亚型在内的肿瘤)的临床数据，针对每个癌型都有规范的，大样本量的临床数据，比在 GEO 数据库中获取的基因表达数据更丰富。TCGA 覆盖了人体六十个组织或器官的三十八种癌型以及亚型，四十个 projects，近四万个患者，而且在不断更新中，更新速度也比其他数据库快，对于癌症的研究有十分丰富的资源。

本文的前列腺基因表达数据来自 TCGA 数据库，在 TCGA 数据库中，样品名称格式为 TCGA-19-2619-10A，其癌症样本和正常样本的分类标准与最后的 10A 有关，编号从 01~09 是癌症样本；10~29 是癌旁样本。按照此分类标准利用 R 软件进行分类筛选，获得前列腺癌与癌旁的基因表达数据，其中包括 488 个患病样本和 12 个健康样本，包括 60,482 条基因(https://cancergenome.nih.gov/)。

2.2. 特征基因提取

通过TCGA数据库得到的基因表达数据为原始数据，要首先对数据进行预处理，利用标准分数(z-score)将得到的前列腺基因表达数据进行标准化处理，得到标准化的数据，其标准化公式为：

$$z_{ij} = \frac{g_{ij} - mean(g_i)}{std(g_i)} \tag{1}$$

其中 g_{ij} 表示基因 i 在样本 j 中的表达值，$mean(g_i)$ 表示基因 i 在所有样本中的平均值，$std(g_i)$ 表示基因 i 在所有样本中的标准差[6]。

传统的特征基因提取方法通常把所有基因 FC (差异表达倍数)的值计算出来，根据 FC 值的正负再筛选出上调基因和下调基因，FC 值为正则为上调基因，FC 值为负则为下调基因。通过此方法提取特征基

因的计算量大，要把六万多条的基因每个都计算一边，耗时多，最后得到的上调基因和下调基因的数量过多，不具有针对性，再拿到 cmap 数据库中去检索，得到的结果也不精确，影响最后的结论。也有研究者采用只求出基因的信息熵和互信息值的方法，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，利用熵和互信息值能有效的提取出复杂疾病的致病基因[7] [8] [9]。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，基因变量 X 的熵表示该模式所包含的信息量公式为：

$$H\left(X\right)=-\sum_{x\in X}p\left(x\right)\log_2^{p(x)} \tag{2}$$

利用公式将得到的数据中 60,482 条基因的信息熵计算出来，取熵值较大即含有较多致病信息的部分基因作为特征基因，再计算这些特征基因之间的互信息值，不仅编程过程复杂且计算量大，而且采取此方法计算得到的互信息矩阵还需要再选取适当的阈值来筛选出可以拿到 cmap 数据库中去比对的基因。而阈值的选取是一个复杂的过程，选的阈值过高或过低都会影响最后的检索结果，所以选取不当会影响结果的精确性，使得检索到的药物不具有代表性和针对性。

针对传统特征提取方法耗时多，计算量大，步骤复杂的情况，本文采取信息熵和差异表达倍数相结合的方法来提取特征基因。首先计算出每条基因的信息熵，然后根据信息熵的值降序排列，取信息熵大即生物学意义高，含有较多致病信息的前 5000 个基因作为特征基因，然后再计算这 5000 个特征基因的基因差异表达倍数值(Fold Change)。前列腺癌的基因表达数据包含 60,482 条基因，如果用上述传统的方法，需要计算 60,482 个基因的 FC 值，采取本方法只需计算信息熵大的 5000 个基因的 FC 值，计算量小，节省了计算时间，FC 值的计算公式为：

$$\log FC_i=\log_2\left(\frac{\frac{1}{S}\sum_{k\in T}z_{ik}}{\frac{1}{M}\sum_{k\in N}z_{ik}}\right) \tag{3}$$

其中 z_{ik} 表示基因 i 在样本 k 中的标准化表达值，T 是癌症病人样本的集合，N 是正常样本的集合，S 和 M 分别为癌症病人样本和正常样本的个数(本文中 $S=488$，$N=12$)，最终将求得的 $\log FC$ 值以 0 为分界点，把 $\log FC$ 值大于 0 的所有基因作为上调基因集合，小于 0 的所有基因筛选为下调基因集合。

2.3. Cmap 分析

Connectivity Map (简称 cmap)是一个基因表达谱数据库，其利用小分子药物处理人类细胞后的基因表现差异，建立一个小分子药物、基因表现与疾病相互关连的生物应用数据库。以基因表达谱为所建立之基因、疾病与药物的关联性，协助学者们在药物开发领域上，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、推论大部分药物分子的主要化学结构，并能够归纳出药物分子可能作用的机制方向。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，有超过 7000 笔的基因表达谱资料。每一种药物分子会以不同浓度(10 nm、100 nm、1 μM 与 10 μM)处理在不同的细胞株(breast、prostate、leukemia 与 melanoma cell line)，并处理不同的时间点(6 与 12 小时)。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[10]。

通过 R 软件筛选出前 5000 个信息熵大的特征基因，并利用公式计算其 FC 值，根据正负将特征基因分成上调基因和下调基因，最后得到上调基因 2778 个和下调基因 2222 个。接下来在上调基因中选取 FC

值较大和下调基因中 FC 值较小的部分基因去检索 cmap 数据库。即将得到的上调基因中降序排列，提取前 200 个 log *FC* 值最大的上调基因作为检索标签，将下调基因升序排列，提取前 200 个 log *FC* 值最小的下调基因作为检索标签，将它们存为.grp 文件，检索 cmap 数据库[11]。将前列腺癌基因表达标签与药物处理基因标签进行统计比较[12]。依据表达谱的相似性给每个前列腺癌-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，对癌症基因具有抑制作用，即可能对前列腺癌具有较好的治疗效果[13] [14]。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[15]。

3. 结果与展望

3.1. 结果分析

Cmap 的分析结果如表 1 所示，根据表中数据可以看出负相关分值最高的是 gliclazide (格列齐特)，分值为−0.69，格列齐特是一种治疗中型糖尿病的药物。经分析和比对，得到的负相关分值最高，表明其对于前列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 Luteolin (木犀草素)、phenoxybenzamine (竹林胺)、alpha-estradiol (α -雌二醇)，naloxone (盐酸纳洛酮)等化合物也具有较高的负相关分值，表明极可能对前列腺癌有治疗效果。其中表中的 alpha-estradiol (α -雌二醇，别名雌二醇)是经皮肤吸收的雌激素治疗剂，目前已经被用来治疗晚期前列腺癌。通过表 1 可以看出排在它上面的药物最后的检索分值的负相关性均高于它，所以这几种药物很可能对治疗前列腺癌有治疗效果。

Table 1. Candidate prostate cancer drugs screened from the connectivity map database
表 1. Connectivity map 数据库筛选出的候选抗前列腺癌药物

cmap name	mean	n	enrichment
gliclazide	−0.69	4	−0.805
lysergol	−0.671	4	−0.748
phenoxybenzamine	−0.661	4	−0.832
abamectin	−0.651	4	−0.675
nipecotic acid	−0.642	4	−0.764
etofenamate	−0.613	5	−0.644
alpha-estradiol	−0.363	16	−0.297

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 为前列腺癌症基因标签与药物基因标签相似的聚合度。

根据在 cmap 数据库中的结果比对显示，gliclazide (格列齐特)极有可能对前列腺癌有治疗作用，作为治疗糖尿病的降糖药物已经成为二型糖尿病一线治疗的标准，并在糖尿病临床治疗中积累了大量成功经验，具有很好的耐受性，多数患者依从性良好。如果有实例进一步验证该药对前列腺癌有治疗作用，那么它作为已经成功上市的药物，则节省了研发新药的时间，对其稍加改善，就可以投入到前列腺癌的治疗中。

3.2. 结论与展望

在提取特征基因过程中，通过表 2 可以看出 re 表示在求得信息熵之后，取前 5000 个熵最大的基因，然后根据公式求得的每个基因的 logFC 值，此方法只需分别计算这 5000 个基因的差异表达倍数即可。简单，易操作，程序编写也容易。但是如果用信息熵和互信息的方法来获取特征基因，如表 3 所示，在同

样求得信息熵之后，取前 5000 个熵最大的基因，根据互信息的定义，要计算出这 5000 个基因两两之间的互信息值，最后得到的 5000 × 5000 的矩阵，编程步骤繁琐，计算量大，而且还要对求得的矩阵选取合适的阈值来筛选特征基因，阈值的选取既不能太高，也不能太低，选取不当会导致特征基因不具有代表性，影响最后到 cmap 数据库检索的结果。

Table 2. The code of using R software to calculate FC value
表 2. 利用 R 软件求 FC 值的代码

```
ZC=read.csv("正常样本.csv"，header='T')
AZ=read.csv("癌症样本.csv"，header='T')
szc=ZC[,c(-1,-2)]
sac=AZ[,c(-1,-2)]
mzc=rowMeans(szc)
mac=rowMeans(sac)
mzc<- matrix(c(1:6),2,3)
mac <- matrix(c(2:7),2,3)
re=log((mac/mzc),2)
```

Table 3. The code of mutual information value using R software
表 3. 利用 R 软件求互信息值的代码

```
nn<-matrix(0,nrow=5000,ncol=5000)
for (i in 1:5000){
for(j in 1:5000){
x <- y.ct.m[i,]
y <- y.ct.m[j,]
nn[i,j]<-mutinformation(x,y)
  }
}
```

本文通过信息熵和 Fold Change 相结合的新方法提取前列腺癌中的特征基因，利用 cmap 数据库将基因与药物进行比对打分，最后得到与治疗前列腺癌有关的药物 gliclazide (格列齐特)、Luteolin (木犀草素)、phenoxybenzamine (竹林胺)等。

与传统的特征基因提取方法比较，本文采取的方法耗时短，在利用 R 软件计算过程中，节省了时间，基于信息熵和 Fold Change 算法提取特征基因，比只利用信息熵和互信息与只利用 FC 值的方法提取特征基因用时少、成本低和准确性高，但数据分析结果还需要临床试验的进一步验证，希望有条件的实验室能完成这一工作。为药物重定位提供了新的途径，推动生物医药产业的发展。

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. (2019) Drug Repositioning for Schizophrenia and Depression/Anxiety Disorders: A Machine Learning Approach Leveraging Expression Data. *IEEE Journal of Biomedical & Health Informatics*, **23**, 1304-1315.

[3] Wang, H., *et al.* (2015) Prediction of Drug-Disease Relations: A Recommendation System Model. *Chinese Pharmacological Bulletin*, **31**, 1770-1774.

[4] Zhang, Y.X., *et al.* (2012) Drug Relocation: An Important Application Field of Cyber Pharmacology. *Chinese Journal of Pharmacology and Toxicology*, **26**, 779-786.

[5] 张晓芳, 康永波, 苏君鸿, 等. 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.

[6] 许凤丹. 基于组织特异性路径与基因突变的肝癌药物重定位研究[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2017.

[7] Zhao, C.H., *et al.* (2017) Building Extraction Method from SVM High Resolution Remote Sensing Image Based on Multi-Feature Fusion. *Journal of Shenyang University* (*Natural Science Edition*), **29**, 314-319.

[8] Ji, X.F., *et al.* (2014) Human Motion Recognition Method Based on AdaBoost Algorithm Feature Extraction. *Journal of Shenyang University of Aeronautics and Astronautics*, **31**, 65-69. https://doi.org/10.1109/CIT.2014.87

[9] 易超, 苏雅婷, 依马木买买提江·阿布拉, 张波, 李海军, 丁伟. Affymetrix 基因表达谱芯片在新疆维吾尔族与汉族胰腺癌组织样本间差异表达基因筛选中的运用[J]. 现代生物医学进展, 2017, 17(32): 6215-6223.

[10] 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.

[11] Ren, L., *et al.* (2017) A Study on the Relocation of Antiradiation Drugs by Comparing the Similarity of Gene Expression Labels. *Journal of Clinical Medicine*, **15**, 19-24.

[12] Wang, K.J., *et al.* (2014) New Opportunities for Drug Research and Development in China: Systematic Drug Relocation Based on Big Pharmaceutical Data. *Chinese Science Bulletin*, **59**, 1790-1796. https://doi.org/10.1360/972013-731

[13] Xiao, S.J., *et al.* (2016) Screening of Hirschsprung's Disease Related Genes and Potential Intervention Molecules Based on Chip Technology and CMAP Database. Southern Medical University, Guangzhou.

[14] Yang, K., Dinasarapu, A.R., Reis, E.S., *et al.* (2013) CMAP: Complement Map Database. *Bioinformatics*, **29**, 1832-1833. https://doi.org/10.1093/bioinformatics/btt269

[15] Steuer, R., Kurths, J., Daub, C.O., *et al.* (2002) The Mutual Information: Detecting and Evaluating Dependencies between Variables. *Bioinformatics*, **18**, S231-S240. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231