徐 畅,王 雪,郭 鑫,等. 基于疾病数据库的文本挖掘工具对比研究基于疾病数据库的文本挖掘工具对比研究[J]. 中华医学图书情报杂志, 2018,27(6):10-15.

DOI:10.3969/j.issn.1671-3982.2018.06.002

· 专题 ·

基于疾病数据库的文本挖掘工具对比研究

徐 畅,王 雪,郭 鑫,李 毅,赵 船,侯跃芳

[摘要]目的:对比分析几种基于疾病数据库的文本挖掘工具,总结各自特点。方法:选择 eRAM、PhenUMA、Gendoo、G2D 4 种工具进行对比分析。以 Rett 综合征为例进行实证研究,筛选与其关联性较高的基因,并结合先验知识做出预测。结果:eRAM、PhenUMA 知识库功能全面,可视化效果好。通过实证研究得到 Rett 综合征相关基因,并结合 PubMed、UniProt 等数据库中的先验知识推测出基因 EGR2、CDKL5、BCHE、DLX5 与 Rett 综合征相关。结论:基于疾病数据库的文本挖掘工具可以有效预测疾病的相关基因,预测疾病、表型、基因间相似和相关关系,有助于疾病的诊断及治疗。

[关键词]疾病数据库;文本挖掘工具;疾病候选基因;比较研究;实证研究

[中图分类号]TP311.13;R-058 [文献标志码]A [文章编号]1671-3982(2018)06-0010-06

Disease databases-based comparison of text mining tools

XU Chang, WANG Xue, GUO Xin, LI Yi, ZHAO Chuan, HOU Yue-fang

(China Medical University Medical Information School, Shenyang 110122, Liaoning Province, China)

Corresponding author: HOU Yue-fang

[**Abstract**] **Objective** To comparatively analyze the disease databases-based text mining tools and their characteristics. **Methods** The 4 text mining tools, namely eRAM, PhenUMA, Gendoo and G2D4, were comparatively analyzed. The high correlation genes in Rett syndrome were screened and predicted according to the priori knowledge in PubMed and UniProt. **Results** The functions and visualization effect of eRAM and PhenUMA were good. The Rett syndrome-related genes were detected in the empirical study which showed that the EGR2, CDK15, BCHF and DLX5 genes were related with Rett syndrome according to the priori knowledge in PubMed and UniProt. **Conclusion** Disease databases-based text mining tools can effectively predict disease-related genes, similarity and correlation among diseases, phenotypes and genes, and can thus contribute to the diagnosis and treatment of diseases.

[**Key words**] Disease database; Text mining tool; Candidate gene of disease; Compareilive stucly; Empirical study

医学研究已进入分子阶段,疾病表型及基因的相似性可能提示分子间的相互作用。由于大多数疾病均为多个基因共同作用的结果,基础医学研究人员通过分子实验确定致病基因的方式费力而耗时,临床研究人员想要针对疾病基因进行治疗也非常困难。新兴的生物信息挖掘技术可以帮助基础医学研究人员在实验前筛选候选基因,也可帮助临床研究人员针对具有相似表型或基因的疾病进行进一步准确诊断治疗及老药新用的尝试[1]。疾病数据库的挖掘对于发现致病基因、阐明分子通路具有重要的

[基金项目]辽宁省教育厅人文社会科学研究项目"基于复杂网络的非相关文献知识发现方法研究"(LR201606);2017 年辽宁省大学生创新创业训练项目"基于疾病遗传数据库的文本挖掘研究"(201710159000031)

[作者单位]中国医科大学医学信息学院,辽宁 沈阳 110122

[作者简介]徐 畅(1996-),女,辽宁本溪人,在读本科生。

[通讯作者]侯跃芳(1972-),女,辽宁海城人,博士,副教授,研究方向为生物医学信息数据挖掘,主持省级课题 3 项,副主编教材 3 部,参编教材 5 部,发文 40 余篇。E-mail:yfhou@cmu.edu.cn

意义,这可以通过疾病表型及基因的相似性比较实现。10 余年来,科研人员开发了多种疾病数据库文本挖掘工具。本文选取 eRAM、PhenUMA 、Gendoo 、G2D 4 种性能良好且运行稳定的免费工具进行对比分析,并利用这些工具进行疾病基因发现的实证研究,力求为疾病的遗传学研究提供准确有效的依据,为临床及基础医学研究人员提供有效的参考信息,提高疾病遗传研究的效率。

1　四种文本挖掘工具

精准医学罕见疾病注释百科全书 eRAM[2](encyclopedia of Rare Disease Annotation for Precision Medicine)(http://www.unimd.org/eRAM/)是由华东师范大学陈庚等人研发的文本挖掘工具。它整合了 10 个知名数据库的疾病数据,主要包括罕见病及其用药门户网站(Orphanet)、人类疾病数据库(MalaCards)、NIH-遗传和罕见疾病(Genetic and Rare Diseases,NGRD)、国际罕见病组织数据库(National Organization for Rare Disorders,NORD),为15 942种罕见疾病提供了丰富的临床和分子注释。在其知识库构建过程中将大量的非结构化数据转化为可操作利用的结构化数据,支持基因、表型、疾病间关系的可视化网络构建。分析结果有准确相关基因和全部相关基因两种参考排序方式,两种分析结果中疾病种类及排序不尽相同,为相关疾病预测提供了更多可能。该工具可用于疾病信息检索、基因型检索、表型网络构建、基因网络构建和疾病对网络构建。eRAM 提供丰富而准确的知识,不仅有助于研究人员探索罕见疾病的潜在机制,而且有助于临床医生做出准确的诊断和治疗决策。

PhenUMA[3](http://www.PhenUMA.uma.es/)是由西班牙马拉加大学 Rocío Rodríguez-López 等人基于生物医学和生物分子数据库中的有效信息建成的独立知识库。它以基因功能和疾病表型关系为基础,构建、分析和可视化生物网络,且分析功能多样化,构建网络可视化效果好。该工具可用于研究功能相关基因之间的新的病理学关系,将疾病归类到特定表型的簇中,发现与表型相关的疾病等。PhenUMA 有助于临床和基础研究人员重新解释其研究结果,并通过优先考虑表面上非相关的隐含因素来重新设计实验。

Gendoo[4](Gene, Disease Features Ontology-based Overview System)(http://Gendoo.dbcls.jp/)由东京大学 Takeru Nakazato 等人研发,通过使用 MeSH 词汇生成相关药物的特征概况、生物现象和解剖结构描述疾病和基因。该工具可用于说明基因和疾病的特征,分别比较基因和疾病特征之间的差异和相似之处,将加速从生物学和临床角度对组学数据的分析。

G2D[5](Genes to Diseases)(http://g2d2.ogic.ca/)由加拿大渥太华健康研究所 Carolina Perez-Iratxeta 等人开发。它通过数据挖掘算法评估疾病映射的染色体区域中的基因优先级。如果表型已经与多个位点连锁,则也可检测来自两个基因位点的蛋白质之间的相互作用。G2D 指出了查询蛋白质和基因组中序列相似性匹配的位置,并利用了现有的关于假基因预测的信息,对识别疾病相关基因具有极大的帮助。

2　研究方法

2.1　4 种工具的对比项目

对比分析 eRAM、PhenUMA、Gendoo、G2D 在运算原理、数据输入、分析功能以及结果输出 4 方面的不同。运算原理的对比项目包括知识库来源、运算方法、创建时间、更新周期,数据输入的对比项目包括可录入数据种类、输入格式,分析功能的对比项目包括功能块、分析起始选项、分析项目,结果输出的对比项目包括输出选项、可视化项目、结果下载格式。

2.2　实证研究

以 Rett 综合征为例,利用上述工具进行疾病基因发现的实证研究。Rett 综合征是一种伴 X 染色体的遗传疾病,多发于女性患者,其发病率为 1/10 000 ~ 1/15 000。患者常表现为脑部发育迟缓、刻板动作、呼吸障碍、运动障碍以及孤独症样的社交障碍,后期可能伴有癫痫的发生[6]。

当前国际权威的在线人类孟德尔遗传数据库(Online Mendelian Inheritance in Man,OMIM)中,查询到 Rett 综合征的相关基因仅有 1 个,为甲基化 CpG 结合蛋白-2(methyl-CpG binding protein 2,MECP2)。因此将 MECP2 作为与 Rett 综合征相关的已知基因。

利用上述 4 种工具对 Rett 综合征的相关基因进行挖掘，选取各工具挖掘结果中得分排位高的前 3 种基因，筛除已知相关基因 MECP2，并去重，进一步验证。

验证方法如下：通过在 PubMed、CNKI、万方等数据库中检索相关文献，验证各工具分析结果中所得基因是否与 Rett 综合征相关；通过在蛋白质数据库 UniProt 中查询某基因的蛋白参与的生物过程和分子功能；通过查阅该生物过程和分子功能是否与 Rett 综合征的病因或症状相关，推断该基因是否可能与 Rett 综合征相关。

3 结果与分析

3.1 运算原理的比较

4 种工具运算原理比较的结果见表 1。

表 1 4 种文本挖掘工具运算原理比较

工具名称	知识库来源	运算方法	创建时间	更新周期
eRAM	OMIM、Orphanet（罕见病及其用药门户网站）、MalaCards、NGRD、NORD、HPO（Human Phenotype Ontology，人类表型本体）、GO（Gene Ontology，基因本体）、UMLS（Unified Medical Language System，一体化医学语言系统）、GWAS（Genome-wide association study，全基因组关联分析）、SNOMED-CT（Systematized Nomenclature of Medicine -- Clinical Terms，医学系统命名法）、MeSH、ICD10（International Classification of Diseases，国际疾病分类）、DOID（Disease Ontology，疾病本体）	夹角余弦法	2017-08	每 6 个月更新
PhenUMA	OMIM、Orphanet、STRING（功能蛋白质关联网络）、Metabolic network（代谢网）、HPO、GO	Resnik 法	2014-07	未更新
Gendoo	OMIM、GO	结果依据 P 值排列，但是具体算法不详	2009-06	2009-09、2010-02、2010-03、2011-01、2012-04 曾有更新
G2D	OMIM、RefSeq（The Reference Sequence database，参考序列数据库）、GO	Resnik 法	2001-06	2002-02、2005-05、2005-11、2005-12、2006-02、2006-03、2006-10、2007-03、2010-10 曾有更新

如表 1 所示，eRAM 整合了来自 13 个数据库的知识，相对完整；Gendoo 和 G2D 知识库来源相对较少。eRAM 是最新创建的，G2D 创建较早，Gendoo 和 G2D 更新次数较多，但在近几年内没有更新。

4 种工具的运算方法总结如下。

eRAM 采用夹角余弦法，通过特征向量对之间的夹角余弦值度量。

PhenUMA 采用 Resnik 法，使用基于 Resnik 方法的两种不同的语义相似性度量计算基因之间的功能相似性和表型谱之间的表型相似性[7]。这两种测量都基于“信息内容”（Information Concept，IC）的概念，它使用每个术语概率（一个术语的注释数与总注释数的比例）的对数进行计算。如果术语的概率降低，则信息内容增加，因此该术语的特异性和信息含量也增加。Resnik 提出，一个给定本体的两个术语之间的语义相似性由最具信息含量的共同祖先（Most Informative Common Ancestor，MICA）的 IC 决定。通过从术语组中所有可能的术语对中选择出最大 MICA 来获取术语组之间的相似性分数。

Gendoo 采用通过比较 OMIM 条目的概况和基因表达数据的聚类结果发现基因组之间的相似性，将所开发的特征概况应用于疾病相关基因的分析，但其具体运算方法不详。

G2D 的运算方法同 PhenUMA。

3.2 数据输入的比较

4 种工具数据输入的比较结果见表 2。

表 2　4 种文本挖掘工具数据输入的比较

工具名称	可录入数据种类	输入格式
eRAM	Gene ID、OMIM ID、Gene Name、Disease Names、UMLS ID、ORPHANET ID	仅可输入单个检索词
PhenUMA	Gene ID(s)、OMIM ID(s)、Gene Names、ORPHANET ID、HPO id	可输入多个检索词
Gendoo	Gene ID(s)、OMIM ID(s)、Gene Names、Disease Names	可输入多个检索词
G2D	Gene ID、MIM number	仅可输入单个检索词

表 2 显示,4 种工具均允许输入基因 ID 号,只有 eRAM、Gendoo 允许输入疾病名称。

eRAM 可录入数据种类最多, G2D 相对可录入数据种类较少。

PhenUMA 和 Gendoo 支持录入多个检索词, eRAM 和 G2D 仅可输入单个检索词。

3.3　分析功能的比较

4 种工具分析功能的比较结果见表 3。

表 3　4 种文本挖掘工具分析功能的比较

工具名称	功能块	分析起始选项	分析项目(相似基因、相似表型)
eRAM	疾病信息检索、基因型检索、表型网络构建、基因网络构建、疾病对网络构建	Precise/fuzzy	disease:gene、phenotype、symptom
PhenUMA	基因网络构建、疾病网络构建、表型网络构建、基因富集分析、疾病富集分析	Network:confidence、output network Enrichment analysis:enrichment type	Phenotypic similarity、functional similarity、inferred relationship by OMIM、inferred relationship by orphan disease、protein interaction、metabolic interaction
Gendoo	基因相关疾病药物检索、疾病相关基因检索、疾病相关药物及生物现象检索	Gene: human、mouse、rat、silkworm、abridopsis、synechocystis、anabaena、mesorhizobium	gene/disease: disease、chemicals、biological phenomena、anatomy、organisms
G2D	基因/表型相关基因检索	Quantity screening	MIM number:gene、disease

4 种工具均支持疾病/表型相关基因检索。eRAM 和 PhenUMA 功能块较多,可分析项目也较多。

eRAM 特色功能块为基因/表型/疾病对的网络构建;PhenUMA 特色功能块为基因/表型/疾病的网络构建和基因/疾病富集分析,且只有 PhenUMA 具有富集分析功能;G2D 功能块较少,但筛选项较完善。

3.4　结果输出的比较

4 种工具结果输出的比较见表 4。

表 4　4 种文本挖掘工具结果输出的比较

工具名称	输出选项	可视化项目	结果下载格式
eRAM	General、phenotype、symptom、genotype、MP(哺乳动物本体)、case 、score	评分列表、两疾病间关联的基因/表型覆盖关系列表、疾病间关系网络图	不可下载,结果能拷贝或截图
PhenUMA	gene、disease、type of relationship、score	评分列表、基因/表型/疾病间关系列表、基因/表型/疾病间关系网络图	不可下载,输出格式为 txt 格式;网络图仅能截图
Gendoo	disease、chemicals、biological phenomena、anatomy、organisms、score	评分列表	不可下载,结果能拷贝
G2D	Function、process、component、score	评分列表	不可下载,结果能拷贝

4 种文本挖掘工具中，eRAM 和 PhenUMA 的分析结果较完全（表 4），可视化效果较好，结果中链接稳定（图 1、图 2）。尤其 PhenUMA 中可给出 4 种表现形式的结果图，且具有筛选功能。G2D、Gendoo 可视化效果相对较差，而且结果中的链接有时失效。

Gendoo 在结果列表中给出与疾病相似度分数，并把分数划分层级，按颜色区分。

G2D 在结果中给出疾病相关的 Mesh 词、Mesh 词出现频率及所在文章和相关基因的本体注释，其结果以列表形式给出。

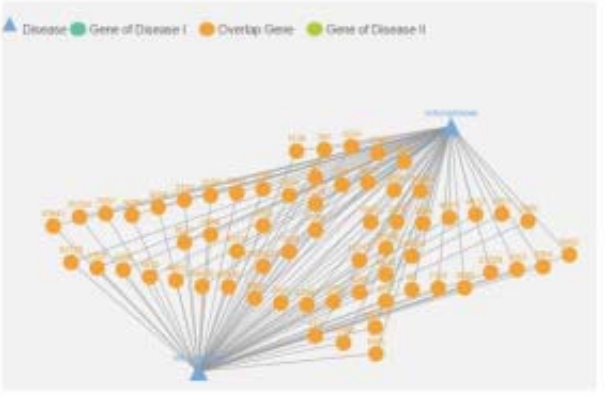

图 1 eRAM 中 Rett 综合征与 schizophrenia 间共享基因网络

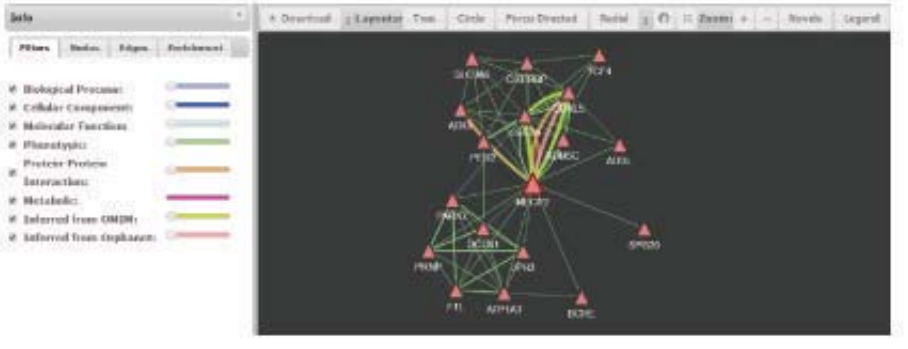

图 2 PhenUMA 中 Rett 综合征的相关基因网络

图 2 的左侧为网络示意图中不同颜色线段的意义说明，可通过调节滑钮筛选结果。

3.5 实证研究结果与分析

OMIM 数据库显示仅有 1 种基因与 Rett 综合征相关。由于 Rett 综合征尚未被认定为单基因疾病，故推测仍有已被认证但未被 OMIM 收录的相关基因，以及未被认证的相关基因。本文利用上述文本挖掘工具可快速发现 Rett 综合征的潜在相关基因。

4 种文本挖掘工具对 Rett 综合征的分析结果如下：eRAM 预测出 6 种相关基因，PhenUMA 在中度置信水平下预测出 21 种相关基因，Gendoo 预测出 56 种相关基因，G2D 预测出 100 种相关基因。选取每种工具分析结果中的前 3 位相关基因（除 MECP2，因 MECP2 是 OMIM 数据库收录的已知 Rett 综合征相关基因）：它们分别是 eRAM 中的 EGR2（early growth response 2）、CDKL5（cyclin-dependent kinase-like 5），PhenUMA 中的 BCHE（butyrylcholin esterase）、CDKL5；Gendoo 中的 CDKL5、DLX5（distal-less homeobox 5；D2D 中的 TAZ（tafazzin）、IKBKG（InhibitorOf Nuclear Factor Kappa B Kinase Subunit Gamma）。

经过去重，得出 6 种相关基因：EGR2、CDKL5、

BCHE、DLX5、TAZ、IKBKG。通过查询 PubMed、UniProt 中关于这些基因的先验知识，进一步分析它们与 Rett 综合征相关的可能性。

EGR2 为序列特异性 DNA 结合转录因子，其参与的脑发育、外周神经系统发育、学习与记忆等生物过程与 Rett 综合征的智力严重低下等症状相关。Swanberg S E 等人研究表明，EGR2 在出生后的人类皮层中发育增加，并在 RTT 和自闭症患者皮质中下调[8]，故推测 EGR2 与 Rett 综合征相关。

CDKL5 介导 MECP2 的磷酸化，可能调控纤毛生成。Vitorino M 等人研究表明 CDKL5 基因突变导致非典型 Rett 综合征[9]。

BCHE 具有广泛底物特异性的酯酶，有助于神经递质乙酰胆碱的失活，可以降解神经毒性有机磷酸酯，其参与的学习、成神经细胞分化、对糖皮质激素的反应等生物过程与 Rett 综合征的智力严重低下、舞蹈样动作、肌张力低等症状相关。

DLX5 涉及骨发育的转录因子，其参与的骨形态发生、口腔发育、头部发育等生物过程与生长迟缓、获得性小头、永久性手的失用、进行性行走困难，步态不稳、躯体的失用和共济失调等症状相关。Proudfoot A 等人研究表明，DLX5 是转录因子，与乳腺癌、肺癌、淋巴瘤、Rett 综合征和人类骨质疏松症有关[10]。

TAZ 的基因编码是在心脏和骨骼肌中高水平表达的蛋白质。经查阅先验知识，推测 TAZ 与 Rett 综合征相关可能性较小。

IKBKG 的基因编码 kappaB 激酶（IKK）是复合物抑制剂的调节亚基，可以激活 NF-κB，导致参与炎症、免疫、细胞存活和其它途径的基因的活化。经查阅先验知识，推测 IKBKG 与 Rett 综合征相关可能性较小。

OMIM 库中仅列 1 种 Rett 综合征相关基因 MECP2。本文利用上述疾病库文本挖掘工具并结合先验知识推测，除 OMIM 数据库所列以外的 4 种相关基因，这是对 OMIM 的有益补充。

4　结论

eRAM、PhenUMA、Gendoo、G2D 4 种工具均可用于快速获取疾病/基因相关信息，并预测疾病与基因的潜在相关关系。eRAM 和 PhenUMA 知识库功能全面，可视化效果好，推荐优先使用。Gendoo 和 G2D 在功能项目上也提供有益的补充，将各工具结合使用可得出更加可信的分析结果。

经实证研究推测基因 EGR2、CDKL5、BCHE、DLX 与 Rett 综合征相关，这可作为 OMIM 数据库对 Rett 综合征相关基因阐述的补充。

基于疾病数据库的文本挖掘工具可以有效预测疾病的相关基因，预测疾病、表型、基因间相似和相关关系，有助于疾病病因及治疗等研究。

【参考文献】

[1] 李建华，李哲人，康　雁，等. 在线孟德尔人类遗传数据库数据挖掘的研究进展[J]. 生物医学工程学杂志，2014(6)：1400-1404.

[2] Jia J, An Z, Ming Y, *et al*. eRAM: encyclopedia of rare disease annotations for precision medicine[J]. Nucleic Acids Research, 2017(1): 937-943.

[3] Rodríguez-López R, Reyes-Palomares A, Sánchez-Jiménez F, *et al*. PhenUMA: a tool for integrating the biomedical relationships among genes and diseases[J]. BMC Bioinformatics, 2014, 15(1): 1-15.

[4] Nakazato T, Bono H, Matsuda H, *et al*. Gendoo: functional profiling of gene and disease features using MeSH vocabulary[J]. Nucleic Acids Research, 2009(37): 166-169.

[5] Pereziratxeta C, Wjst M, Bork P, *et al*. G2D: a tool for mining genes associated with disease[J]. Bmc Genetics, 2005, 6(1): 1-9.

[6] 张晶晶，包新华. Rett 综合征的致病基因彻 CECP2 的研究进展：MECP2 的基因结构、功能及调控基因[J]. 北京大学学报：医学版，2009，41(6)：712-715.

[7] 邵玉凯. 基于人类表型本体的基因和疾病关联关系分析方法研究[D]. 哈尔滨：哈尔滨工业大学，2015.

[8] Swanberg SE, Nagarajan RP, Peddada S, *et al*. Reciprocal co-regulation of EGR2 and MECP2 is disrupted in Rett syndrome and autism[J]. Human Molecular Genetics, 2009, 18(3): 525.

[9] Vitorino M, Cunha N, Conceião N, *et al*. Expression pattern of CDKL5 during zebrafish early development: implications for use as model for atypical Rett syndrome[J]. Molecular Biology Reports, 2018, 45(8): 445-451.

[10] Proudfoot A, Axelrod HL, Geralt M, *et al*. Dlx5 homedomain/DNA complex: Structure, binding and effect of mutations related to split-hand and foot malformation syndrome[J]. Journal of Molecular Biology, 2016, 428(6): 1130-1141.

[收稿日期：2018-05-05]

[本文编辑：黄思敏]