

评测纵览：面向“基因 - 疾病”的关联语义挖掘任务*

欧阳思卓 姚昕智 王宇星 彭钱钱 贺芷涵 夏静波

(1 华中农业大学信息学院 武汉 430070

2 华中农业大学农业生物信息学湖北省重点实验室 武汉 430070)

〔摘要〕 介绍活跃基因注释语料库的设计目的和标注过程、特点，阐述标注规则、标签设计、标注实例，分析语料库在医药领域的应用场景，包括抗癫痫药物重定位、阿尔茨海默症关键基因提取等领域。

〔关键词〕 语料库注释；药物重定位；癌症基因预测；生物过程提取；人工智能

〔中图分类号〕 R - 058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.12.002

An Overview of the Text Mining Task for “Gene - Disease” Association Semantics OUYANG Sizhuo, YAO Xinzhi, WANG Yuxing, PENG Qianqian, HE Zhihan, XIA Jingbo, 1College of Informatics, Huazhong Agricultural University, Wuhan 430070, 2Hubei Key Lab of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China

〔Abstract〕 The paper introduces the design purpose, annotation process and characteristics of the Active Gene Annotation Corpus (AGAC), expounds annotation rules, label design and annotation examples, and analyzes the application scenarios of the corpus in the field of medicine, including the repurposing of antiepileptic drugs, the extraction of key genes of Alzheimer’s disease, etc.

〔Keywords〕 corpus annotation; drug repurposing; cancer gene prediction; biological process extraction; Artificial Intelligence (AI)

1 引言

1.1 AGAC 语料库设计目的和标注过程

1.1.1 设计目的 活跃基因注释语料库 (Active

〔修回日期〕 2022 - 09 - 23

〔作者简介〕 欧阳思卓，硕士研究生，发表论文 1 篇；通信作者：夏静波，博士，副教授，发表论文 90 篇。

〔基金项目〕 中央高校基本科研业务费专项资金“人工智能共性平台的理论与应用研究”（项目编号：2662021JC008）；中央高校基本科研业务费专项资金“针对稻、麦、玉米三类主要粮食作物表型挖掘的语料资源构建”（项目编号：2662022XXQD001）。

Gene Annotation Corpus, AGAC) 是针对具有改变中心功能变化的基因标注问题设计的^[1]。“功能变化”的定义源自于对功能丧失 (Loss of Function, LOF) 和功能获得 (Gain of Function, GOF) 的聚焦和一个直观的药理学假设^[2]：如果一个基因突变后功能丧失或者功能获得引起某种疾病，则针对 LOF/GOF 突变基因的拮抗剂或者激动剂很有可能成为该疾病的特效药物。

1.1.2 语料库标注 AGAC 语料库由 1 名主标注员和 3 名副标注员历时 17 个月标注完成，语料库中包含了从 PubMed 收集的 500 篇人工注释摘要，挑选了 8 个触发词实体，涵盖从分子水平到细胞水平的生物现象和过程，其中包括 5 类生物概念实体（突变、相互作用、通路、分子生理活性、细胞生

理活性) 和 3 类调控概念实体 (正调控、负调控和调控)。此外, AGAC 通过主事和致事两个语义角色描述主题和因果关系。

1.1.3 语料库注释质量检验 AGAC 语料库通过标注者间信度检验注释质量。在包含严格边界检查的严苛比较标准下, 平行标注结果精确率均高于召回率。针对出现的标注异议, 主标注员依据指南^[3]进行裁议。因此, 综合测试结果和标注流程控制, 语料库标注质量得到有效保障。

1.2 AGAC 语料库特点

1.2.1 数据不平衡 当对所有文本标注情况进行统计时, 统计数据表明标注出的实体分布情况在类型上是不平衡的。例如“突变 (Var)” 标签出现了上千次但是“通路 (Pathway)” 标签总共只出现了几十次。语义角色的标注中也出现此类分布不平衡现象, “CouseOf” 关系的注释次数在测试集和训练集相差较大。

1.2.2 选择性注释 根据 AGAC 的标注指南^[3], 标注时并不是无差别注释, 而是根据标注规则选择对含有足够所需信息的句子进行注释, 以期挖掘出带有明确 LOF/GOF 描述的“基因 - 疾病” 关联的

病理信息。这导致基于 AGAC 语料库的命名实体识别具有挑战性, 所设计算法需在语义信息提取上有深入理解。

1.2.3 潜在主题注释 通过文本挑选规则选取的每一个带有 LOF/GOF 类别的“基因 - 疾病” 关联摘要都会被认为是一个潜在的主题注释, 这一特性将 AGAC 与其他语料库区别开来, 使其注释具有独特性, 真正面向药理学假设 LOF 拮抗剂和 GOF 激动剂的知识发现, 用于药物重定位和药物再利用的场景^[4]。

2 AGAC 语料库标注指南

2.1 标注规则和标签设计

AGAC 触发词实体标注规则遵循生物学的中心法则, 采用选择性标注, 语义不完整的句子不予标注。标注规则^[3]确保逻辑链条起点是“突变” 等变异信息, 利用分子事件来聚焦分子生理活性, 记录实体间相互作用, 筛选生物通路信息。语料库实体标签说明和实例, 见表 1。另外, 两个语义角色标注提供了实体之间关系的表征, “ThemeOf” 用来描述主事实体指向当前实体的关系, “CauseOf” 用来描述当前实体指向致事实体的关系。

表 1 触发词实体标签说明

实体标签	英文全称	中文全称	标签说明	标注实体实例
Var	Variation	突变	DNA、RNA、蛋白质的突变和分子结构的变化	“M489V mutation”, “germline mutations”
Interaction	Interaction	相互作用	分子、细胞之间的相互影响	“bind”, “interaction”
Pathway	Pathway	通路	信号转导和细胞代谢等通路	“Bmp pathway”, “PI3K pathway”
MPA	Molecular Physiological Activity	分子生理活性	分子水平的活性	“phosphorylation”, “nigral vulnerability”
CPA	Cellular Physiological Activity	细胞生理活性	细胞水平或以上的活动	“calvarial osteoblasts”, “osteoblast proliferation”
PosReg	Positive Regulation	正调控	获得功能的线索词	“facilitates”, “increased”
NegReg	Negative Regulation	负调控	失去功能的线索词	“suppressed”, “decreased”
Reg	Regulation	调控	中性的线索词	“resulted in”, “regulated”

2.2 实体标注和语义标注示例

根据设定的标注规则, 标注示例, 见图 1。该句子中有 5 个实体被识别出来, 其中, “TRPV1” 是一个基因, 被标注为 “Gene”; “gain - of - function” 获

得功能是一个正调控的实体, 被标注为 “PosReg”; “mutation” 被标注为 “Var”; 而谓语动词 “impairs” 是一个表示负向调控的实体词, 被标注为 “NegReg”; “pain and itch sensations” 疼痛和瘙痒是发生在细胞层面的活动, 因此被标注为 “CPA”。语义角

色也被识别出 4 个，其中，R1 是由“TRPV1”指向“mutation”的“ThemeOf”关系，因为该突变发生的主体是“TRPV1”基因；R2 是由“mutation”指向“gain-of-function”的“CauseOf”关系，因为突变发生后导致了 TRPV1 基因获得功能；R3 是由“gain

- of - function”指向“impairs”的“CauseOf”关系，因为 TRPV1 基因突变之后获得功能导致后续其他生物活动减弱；R4 是由“pain and itch sensations”指向“impairs”的“ThemeOf”关系，因为“减弱”事件的主体是“pain and itch sensations”。

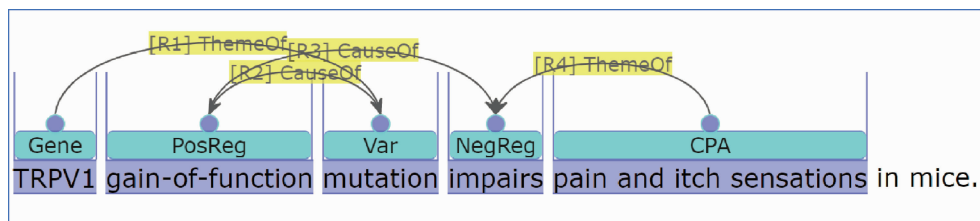


图 1 标注示例（原文来自 PMID: 29424270）

3 AGAC 语料库在医药领域应用场景

3.1 概述

根据 AGAC 语料库标注后的结果应用于实际场景获得的启发，发现其与与癌症有关的关键基因预

测、药物重定位和疾病病理循证等场景下都取得一定的成效，通过预测基因和药物对成功找到有文献支持的 6 种可能有抗癫痫疾病作用的药物，通过对基因突变之后失去功能或者得到功能的预测可以提取导致阿尔茨海默症的关键基因，见图 2。

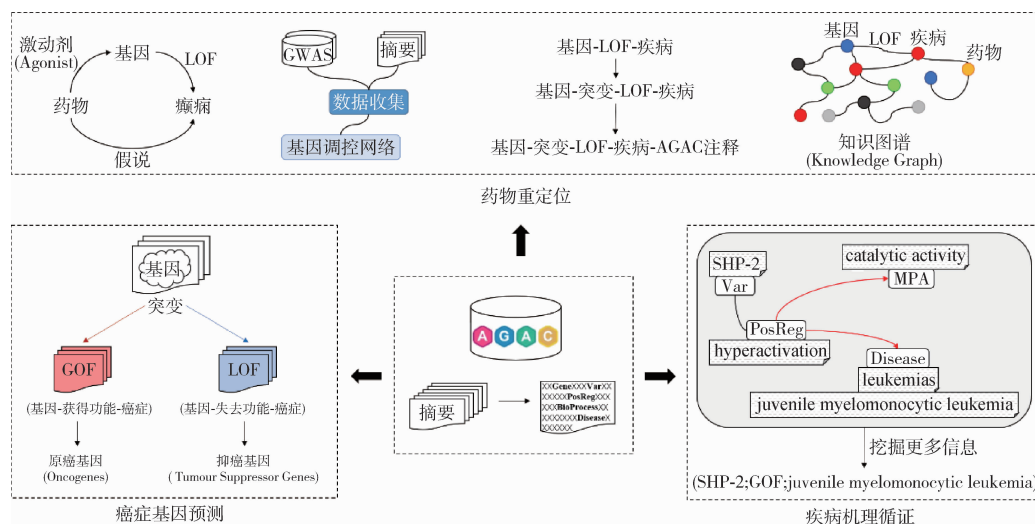


图 2 AGAC 语料库在关键基因预测、药物重定位知识发现场景的应用

3.2 AGAC 应用于阿尔茨海默症中关键基因的预测

3.2.1 应用价值 阿尔茨海默症（Alzheimer's Disease, AD）是一种常见的神经退行性疾病，会损害记忆力、语言和各种身体行为。虽然没有数据库记录 AD 相关基因的突变类型（LOF 突变/ GOF 突变）信息，但有大量文献报道了 AD 发病机制的研究情况。因此，AGAC 技术可以被应用于 AD 文献

中，以提取突变基因及其改变的生物学过程^[5]。

3.2.2 应用方法 首先在 PubMed 上获取 AD 相关的文献，基于规则和文本相关性进行过滤，将过滤后的 AD 文本进行 AGAC 标注，并推断文本中包括的 LOF/GOF 突变信息，共提取出 325 个突变并得到了 822 对 LOF/GOF 相关的突变三元组信息及其对应的句子证据。这 325 个突变带有突变后下游生物过程的清晰语义，经过人工整理后可分为 8 种

类型。包括 5 类实体信息：基因表达、蛋白质活性、相互作用、通路活性和细胞活性，这 5 类都是遵循了中心法则从分子水平到细胞水平的基本生物学过程。除此之外还包括磷酸化、A β 积累和钙离子浓度 3 类语义信息。而相关的句子证据可以证明提取出的三元组信息的正确性，例如来自 PubMed 唯一标识码（PubMed Unique Identifier, PMID）为 29656768 的文献摘要中的一个三元组是“MLKL; LOF; Alzheimer's disease”，它对应的句子证据是“MLKL loss - of - function mutation might contribute to late - onset ApoE ϵ 4 - negative AD in the Hong Kong Chinese population.”。从生物学的观点来看，失去功能（loss - of - function）显然是功能失去的一种描述。因此，这句话承载着明确的语义信息，即基因“MLKL”在突变后发挥着与“阿尔茨海默症”相关的“LOF”功能。

3.3 AGAC 在抗癫痫药物重定位中的应用

通过在实际场景中的应用得知，使用 AGAC 可以优化抗癫痫药物的发现过程^[1]，通过全基因组关联分析（Genome - wide Association Study, GWAS）等多个数据集收集与癫痫相关的基因作为关键词在 PubMed 数据库检索文献并提取摘要，再对这些文本以“突变”和“癫痫”为关键字过滤后进行自动化标注，通过 AGAC 获得功能和失去功能规则对标注后的句子进行分类，根据前文的药理学假说将预测结果与 DrugBank 数据库中条目匹配，获得了 281 个“基因 - 药物”对，包括了 112 种药物和 28 个基因，112 种药物中有 30 种药物被录入数据库。新预测出的 10 种未入库的多靶点药物中，有 6 种被发现与癫痫相关并且都得到文献支持，这 6 种药物分别是：奥沙西泮、替马西泮、哈拉西泮、普拉西泮、唑吡坦和硫戊巴比妥。该案例研究结果支持了 AGAC 在促进知识发现和药物再利用方面的潜力。

4 结语

本文基于 AGAC 语料库设计了一个面向“基因 - 疾病”关联的语义挖掘任务，并简要介绍在构建语料库时设计的文本收集规则、实体标注和语义角色标注规则以及语料库“数据不平衡”“选择性注释”和“潜在主题注释”3 个特点，提出语料库在抗癫痫药物重定位、提取阿尔茨海默症关键基因等领域的应用，体现出 AGAC 语料库应用前景和潜力，促进出物自然语言处理和医学健康等相关方向的跨学科应用。

参考文献

- 1 Wang Y, Zhou K, Kim J D, et al. An Active Gene Annotation Corpus and Its Application on Anti - epilepsy Drug Discovery [C]. San Diego: The IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019.
- 2 Wang Z Y, Zhang H Y. Rational Drug Repositioning by Medical Genetics [J]. Nature Biotechnology, 2013, 31 (12): 1080 - 1082.
- 3 Wang Y, Yao X, Zhou K, et al. Guideline Design of an Active Gene Annotation Corpus for the Purpose of Drug Repurposing [C]. Beijing: 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP - BMEI 2018), 2018.
- 4 Wang Y, Zhou K, Gachloo M, et al. An Overview of the Active Gene Annotation Corpus and the BioNLP OST 2019 AGAC Track Tasks [C]. Hong Kong: The 5th Workshop on BioNLP Open Shared Tasks, 2019.
- 5 Zhou K, Wang Y, Bretonnel Cohen K, et al. Bridging Heterogeneous Mutation Data to Enhance Disease Gene Discovery [EB/OL]. [2022 - 03 - 15]. https://www.researchgate.net/publication/350872250_Bridging_heterogeneous_mutation_data_to_enhance_disease_gene_discovery.