基于语义关系发现的阿尔茨海默病药物重定位*

薛 均 施 维 潘璀然 王青华 王 理 蒋 葵 董建成

(南通大学医学院医学信息学系 南通 226001)

〔摘要〕 针对阿尔茨海默病，提出基于语义关系 SCP 算法发现重定位药物的方法，从总体流程、语义关系集抽取等方面阐述算法原理，介绍结果验证方法并比较该算法与 LTC－AMW 算法筛选药物－AD 关系对排序结果。

〔关键词〕 药物重定位；语义关系；SemMedDB；阿尔茨海默病

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.04.013

Relocation of Alzheimer's Disease Drugs That Is Discovered Based on Semantic Relation *XUE Jun, SHI Wei, PAN Cui－ran, WANG Qing－hua, WANG Li, JIANG Kui, DONG Jian－cheng, Department of Medical Informatics of Medical School, Nantong University, Nantong 226001, China*

〔**Abstract**〕 Aiming at Alzheimer's disease, the paper puts forward the method of relocating drugs, which is discovered based on semantic relation SCP algorithm; it expounds on the algorithm principle from aspects like overall process and semantic relation set extraction, introduces result verification method and compares the drugs screened through the algorithm and LTC－AMW algorithm, the heapsort result of AD relation.

〔**Keywords**〕 Drug relocating; Semantic relation; SemMedDB; Alzheimer's disease

1 引言

目前新药研发费用平均 20～30 亿美元，研发周期大约需要 13～15 年时间[1]。新药研发费用正逐年增加，上市新药却不断减少，新药研发效率正逐年降低。为降低药物的研发周期、成本和风险，药物重定位逐渐成为药物研发的重要策略。据估计重定位药物费用平均为 3 亿美元，研发周期约 6.5 年，越来越受到科研机构、医药企业的广泛关注。近年来重定位的药物约占美国食品药品管理局(Food and Drug Administration，FDA)批准药物和疫苗的 30%[2]。

Semantic Medline Database (SemMedDB) 包含大约 7 000 万个语义关系[3]，是使用基于规则的自然语言处理系统 SemRep[4]，从所有 Medline 标题和摘要中提取实体之间语义关系。Zhang R、Cairelli M J 和 Fiszman M 等人使用 SemMedDB 中的语义关系，

〔修回日期〕 2017－12－15

〔作者简介〕 薛均，硕士；通讯作者：王理，副教授。

〔基金项目〕 国家自然科学基金项目“糖尿病信息管理系统中视网膜图像互操作与 CAD－SR 研究”(项目编号：81501559)；国家自然基金项目“电子健康档案系统中临床医生信息的集成与可视化研究”(项目编号：81701793)；江苏省研究生科研与实践创新计划项目“中文电子病历命名实体识别”(项目编号：KYCX17－1932)。

通过药物 - 基因 - 癌症、药物 - 基因 - 基因 - 癌症路径模式用于潜在的前列腺癌药物的发现[5]。Yang H T、Ju J H 和 Wong Y T 等人通过大规模的文献挖掘，从 SemMedDB 中检索所有药物 - 基因、基因 - 疾病的语义关系，发现潜在治疗疾病的重定位药物[6]。如何从大量尚未验证的药物和疾病关系中发现具有潜在治疗作用的药物 - 疾病关系是药物重定位研究的关键。借助自然语言处理技术和大量可用的医学数据资源，如生物医学文献，电子病历，医疗数据库等，可以显著降低成本来鉴定和验证药物重定位候选者。汪浩，王海平，吴信东等人将社交网络中推荐模型应用于重定位研究，预测具有潜在治疗关系的药物 - 疾病[7]。叶浩、杨琳琳和曹志伟等人基于 pahway 谱的药物 - 疾病关联评估方法来预测药物新的治疗适应症[8]。林耀进、张佳和林梦雷等人通过收集药物及疾病的信息构建药物 - 疾病关联矩阵，基于协同过滤的药物重定位算法预测对某特定疾病有治疗作用的药物 - 疾病组合[9]。

2 算法原理

2.1 总体流程

华盛顿大学 Swanson 教授最早提出生物医学文献中可能隐藏着大量不为人知的科学知识这一假设，由此 Swanson 提出非相关文献知识发现的模型：ABC 模型[10-11]。根据 ABC 模型，提出基于语义关系的 SCP 算法发现重定位药物的方法，具体实现流程，见图 1。

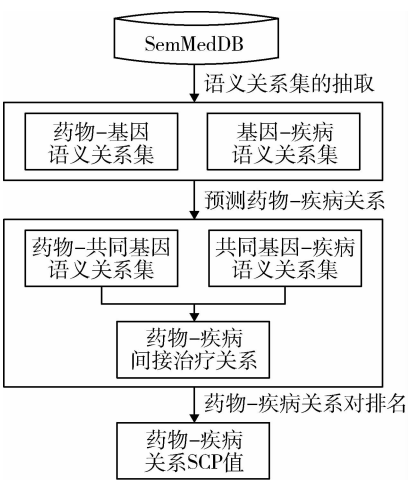

图 1 基于语义关系的药物重定位总体流程

2.2 语义关系集抽取

2.2.1 关系谓词确定 从 SemMedDB 中抽取基因和疾病以及药物和基因之间的语义关系集，根据 Ahlers C B，Fiszman M 和 Demner - Fushman D 等人的研究，在抽取基因和疾病语义关系集时，关系谓词限定为以下 6 种类型：AFFECTS、ASSOCIATED_ WITH、AUGMENTS、CAUSES、DISRUPTS、PRE DISPOSES[12]；在抽取药物和基因语义关系集时，关系谓词限定为以下 3 种类型：STIMULATES、INHIBITS、INTERACTS_ WITH[12]。

2.2.2 语义类型抽取 根据医学一体化语言系统（Unified Medical Language System，UMLS）语义网络中语义类型[13]，选取与药物、基因、疾病相关的语义类型。实体语义类型，见表 1。此外，DrugBank 是世界上最大、最完整的药物数据库，因此选取的药物均为 DrugBank 数据库中出现的药物，基因参照 HUGO 数据库匹配到标准的基因名称缩写[14]。

表 1 实体语义类型

实体	语义类型缩写	语义类型全称
药物	aapp	氨基酸、多肽和蛋白质（amino acid, peptide, or protein）
	antb	抗生素（antibiotic）
	clnd	临床药物（clinical drug）
	horm	激素（hormone）
	imft	免疫因子（immunologic factor）
	nnon	核酸、核苷、核苷酸（nucleic acid, nucleoside, or nucleotide）
	opco	有机磷化物（organophosphorus compound）
	orch	有机化学物质（organic chemical）
	phsu	药物学物质（pharmacologic substance）
基因	aapp	氨基酸、多肽和蛋白质（amino acid, peptide, or protein）
	gngm	基因或基因组（gene or genome）
疾病	dsyn	疾病或综合征（disease or syndrome）
	mobd	心理或行为异常（mental or behavioral dysfunction）
	neop	病理过程（neoplastic process）

2.3 预测药物 – 疾病关系对

药物与疾病通过共同基因相关联，如药物二甲双胍（Metformin）可以作用于 A2M、INS 和 MAOB 等基因，A2M、INS 和 MAOB 等基因同时与阿尔茨海默症（Alzheimer′s Disease，AD）都有关联。针对药物二甲双胍和阿尔茨海默病，共同基因为 A2M、INS 和 MAOB 等基因。本文针对 AD 展开研究，从 SemMedDB 中抽取药物 – 基因、基因 – 疾病的语义关系集，推断潜在的药物 – AD 治疗关系。如从例 1、例 2 的语义关系中可以预测 Selegiline – Alzheimer's Disease 为潜在药物 – 疾病关系对。

例 1：Selegiline | orch，*INTHIBITS*，MAOB | gngm

实体 1 | 语义类型，关系谓词，实体 2 | 语义类型

例 2：MAOB | gngm，*AUGMENTS*，Alzheimer's Disease | dsyn

实体 1 | 语义类型，关系谓词，实体 2 | 语义类型

2.4 基于 SCP 算法的药物 – AD 关系对排名

使用对称条件概率（Symmetrical Conditional Probility，SCP）计算 $W_{G(g_i,d_j)}$，如公式（1）。

$$W_{G(g_i,d_j)} = \sqrt{P(g_i \mid d_j)P(d_j \mid g_i)} \tag{1}$$

公式（1）主要目的是通过相互条件概率计算基因 g_i 和疾病 d_j 的关联程度。

$$W_{G(g_i,d_j)} = \sqrt{\frac{f(g_i,d_j)^2}{f(g_i)\cdot f(d_j)}} \tag{2}$$

其中 G 代表基因且 $g_i \in G$，D 代表疾病且 $d_j \in D$，公式（2）计算基因和疾病之间的关联权重。$f(g_i, d_j)$ 为基因 g_i 与疾病 d_j 共同出现的次数，$f(g_i)$ 为基因 g_i 总共出现的次数，$f(d_j)$ 为疾病 d_j 总共出现的次数。

$$W_{T(t_i,g_j)} = \sqrt{\frac{f(t_i,g_j)^2}{f(t_i)\cdot f(g_j)}} \tag{3}$$

其中 T 代表药物且 $t_i \in T$，公式（3）计算药物和基因之间的关联权重。$f(t_i, g_j)$ 为药物 t_i 与基因 g_j 共同出现的次数，$f(t_i)$ 为药物 t_i 总共出现的次数，$f(g_j)$ 为基因 g_j 总共出现的次数。

$$SCP(t,d) = \sqrt{\frac{f(t,C)}{f(t)}\cdot\frac{f(C,d)}{f(d)}} \tag{4}$$

其中 C 代表连接药物 t 和疾病 d 的共同基因，共同基因由连接药物 t_i 与疾病 d_j 的基因组合构成。$SCP(t,d)$ 表示药物 t 和疾病 d 的关联程度。

$$f_T(t,C) = \sum_{g\in C} W_T(t,g) \tag{5}$$

其中 g 代表基因，$f_T(t,C)$ 代表药物 t 与共同基因 C 的权重之和。

$$f_D(C,d) = \sum_{g\in C} W_D(g,d) \tag{6}$$

其中 g 代表基因，$f_D(C,d)$ 代表共同基因 C 与疾病 d 的权重之和。

$$f(t) = \sum_{g\in G} W_T(t,g) \tag{7}$$

其中 G 代表药物 – 基因语义关系集中出现的与药物 t 相连的全部基因。$f(t)$ 代表语义关系集中药物 t 和基因 G 的权重之和。

$$f(d) = \sum_{g\in G'} W_D(g,d) \tag{8}$$

其中 G′代表基因 – 疾病语义关系集中出现的与疾病 d 相连的全部基因。$f(d)$ 代表语义关系集中基因 G′和疾病 d 的权重之和。

3 结果

3.1 验证方法

为验证本研究所提出的基于语义关系的药物重定位方法，通过公式（2）计算得出与阿尔茨海默病相关联的基因权重排序。通过参照 Genetics Home Reference（GHR）和 Online Mendelian Inheritance in Man（OMIM）中的基因文献记录进行标注，在得到的前 50 个基因中有 39 个与 AD 有关联关系。实验结果为提取与 AD 有关的基因的准确率为 78%。通过公式（3）计算药物与共同基因的权重，通过公式（4）得出药物与 AD 的间接关系的 SCP 值，根据 SCP 值将与 AD 相关的药物进行排序。为验证预测药物 – AD 关系对的有效性，将 CTD 提供的已知药物与疾病关系对作为金标准[16]，预测的药物 – AD 关系对有 62.2% 得到 CTD 已知的药物与疾病关系对支持。

3.2 两种算法筛选药物 – AD 关系对结果比较

Yetisgen – Yildiz 和 Prattp 提出中间链接词数量的平均最小权重（Linking Term Count with Average Minimum Weight，LTC—AMW）在评估目标词排序

中有良好的性能[16]。根据初始词和中间词，中间词和目标词的平均最小 MIM 权重计算目标词排序。本研究中疾病为初始词，基因为中间词，药物为目标词。为验证 SCP 算法计算药物 - AD 关系对排序方法，与 LTC - AMW 计算的药物与 AD 关系对排序结果比较。两种算法筛选药物 - AD 关系比较，见图 2。基于语义关系 SCP 算法得出的排序前 100 名的药物 - AD 关系对中有 82% 出现在 CTD 数据库中，而 LTC - AMW 算法得出的排序前 100 名的药物 - AD 关系对中 76% 在 CTD 数据库中出现。比较两种算法，SCP 算法识别排名前 100 名药物 - 疾病关系对的准确度高于 LTC - AMW 算法。但对于排名 100 之后的药物 - 疾病关系对，SCP 算法的结果准确度接近 LTC - AMW 算法。通过 SCP 算法筛选出排名较高的药物 - AD 关系对，可以明显提高潜在治疗 AD 重定位药物的富集程度。基于语义关系的 SCP 算法对排名前 10 名的药物通过文献查证进行人工评估分析[17-24]，具体结果，见图 3。其中 8 对通过文献查证得到临床实验和综述报道的支持。其中 may - treat 表示药物与 AD 可能存在治疗关系，induce 表示药物诱发产生 AD 样改变，unknown 表示药物与 AD 未知或者不存在治疗关系。

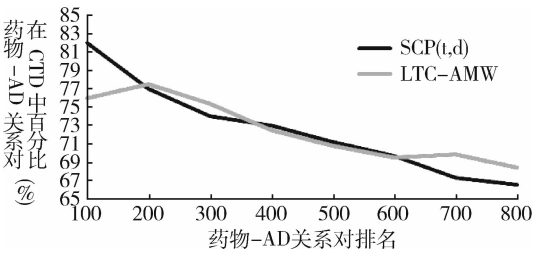

图 2 两种算法筛选药物 - AD 关系比较

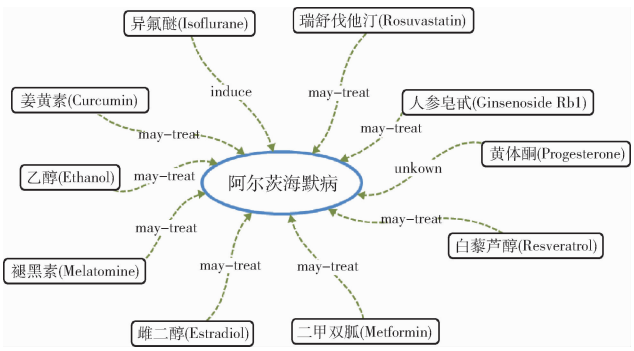

图 3 与阿尔茨海默病关联的排名前 10 的药物

4 结语

基于语义关系的 SCP 算法筛选治疗阿尔茨海默病的重定位药物，缩小筛选阿尔茨海默病重定位药物的范围，促进对潜在的阿尔茨海默病治疗药物的发现。评估预测药物 - 疾病关系出现在 CTD 中的百分比，排名高的药物更有可能是治疗疾病重定位药物的候选者。在算法方面，由于预测的药物 - 疾病关系对并没有考虑药物和疾病关系对具体的语义关系（may - treat 或者 induce），结果的准确率有所欠缺，下一步将研究如何从药物 - 基因语义关系和基因 - 疾病语义关系推断药物 - 疾病语义关系，筛选药物 - 疾病关系对中存在可能治疗关系的药物，提高药物重定位的准确性，降低预测的假阳性率。

参考文献

1 Nosengo N. Can You Teach Old Drugs New Tricks? [J]. Nature, 2016, 534 (7607): 314 - 316.

2 Jin G, Wong S T. Toward Better Drug Repositioning: prioritizing and integrating existing methods into efficient pipelines [J]. Drug Discovery Today, 2014, 19 (5): 637 - 644.

3 Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed - scale repository of biomedical semantic predications [J]. Bioinformatics, 2012, 28 (23): 3158 - 3160.

4 Rindflesch T C, Kilicoglu H, Fiszman M, et al. Semantic MEDLINE: an advanced information management application for biomedicine [J]. Information Services & Use, 2011, 31 (1 - 2): 15 - 21.

5 Zhang R, Cairelli M J, Fiszman M, et al. Exploiting Literature - derived Knowledge and Semantics to Identify Potential Prostate Cancer Drugs [J]. Cancer Informatics, 2014, 13 (Suppl 1): 103 - 111.

6 Yang H T, Ju J H, Wong Y T, et al. Literature - based Discovery of New Candidates for Drug Repurposing [J]. Briefings in Bioinformatics, 2016, 18 (3): 488 - 497.

7 汪浩，王海平，吴信东，等. 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，(12)：1770 - 1774.

8 叶浩，杨琳琳，曹志伟，等. 基于 pathway 谱对药物重定

位的探讨［J］. 科学通报, 2012, 57（7）: 534 - 541.
9 林耀进, 张佳, 林梦雷, 等. 基于协同过滤的药物重定位算法［J］. 南京大学学报（自然科学）, 2015, 51（4）: 834 - 841.
10 Swanson D R. Undiscovered Public Knowledge［J］. Library Quarterly, 1986, 56（2）: 103 - 118.
11 Swanson D R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge［J］. Perspectives in Biology & Medicine, 1986, 30（1）: 7 - 18.
12 Ahlers C B, Fiszman M, Demner - Fushman D, et al. Extracting Semantic Predications from Medline Citations for Pharmacogenomics［J］. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2007,（12）: 209 - 220.
13 U. S. National Library of Medcine. Current Semantic Types.［EB/OL］.［2017 - 11 - 01］. https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.
14 Yates B, Braschi B, Gray K A, et al. Genenames. org: the HGNC and VGNC resources in 2017［J］. Nucleic Acids Research, 2016, 41（21）: 9680 - 9687.
15 Davis A P, Grondin CJ, Johnson RJ, et al. The Comparative Toxicogenomics Database: update 2017［J］. Nucleic Acids Research, 2017, 45（D1）: D972 - D978
16 Yetisgen - Yildiz M, Pratt W. A New Evaluation Methodology for Literature - based Discovery Systems［J］. Journal of Biomedical Informatics, 2009, 42（4）: 633 - 643.
17 Shuai Z, Hu X, Wei G, et al. Isoflurane Anesthesia Promotes Cognitive Impairment by Inducing Expression of β - amyloid Protein - related Factors in the Hippocampus of Aged Rats［J］. Plos One, 2017, 12（4）: 1 - 14.
18 Berntsen S, Kragstrup J, Siersma V, et al. Alcohol Consumption and Mortality in Patients with Mild Alzheimer's Disease: a prospective cohort study［J］. BMJ Open, 2015, 5（12）: 1 - 7.
19 Cardinali D P, Brusco L I, Liberczuk C, et al. The Use of Melatonin in Alzheimer's disease［J］. Neuro Endocrinology Letters, 2002, 23（1）: 20 - 23.
20 杨宇, 梁梅冰, 贾真, 等. 姜黄素在阿尔茨海默病中对炎症以及神经元的保护机制研究［J］. 武汉大学学报（医学版）, 2015, 36（3）: 332 - 336.
21 饶艳秋, 王文君. 雌激素防治阿尔茨海默病的作用机制［J］. 国际妇产科学杂志, 2014, 41（1）: 32 - 34.
22 钱钧强, 叶因涛, 王冬, 等. 白藜芦醇治疗阿尔茨海默病的研究进展［J］. 现代药物与临床, 2016, 31（6）: 924 - 928.
23 Zissimopoulos J M, Barthold D, Brinton R D, et al. Sex and Race Differences in the Association Between Statin Use and the Incidence of Alzheimer Disease［J］. Jama Neurology, 2017, 74（2）: 225 - 232.
24 李菁媛, 李乃静. 人参皂苷治疗阿尔茨海默病的药理作用及机制研究进展［J］. 实用老年医学, 2017,（7）: 606 - 608.

（上接第 53 页）

验。技术的不断发展给予持续改进、推陈出新的动力，应不断地探索充分利用新技术的优势促进工作更高效、便捷。

参考文献

1 黄爱仪. 浅析高校机房管理办法的创新和探索［J］. 科技信息, 2010,（3）: 53 - 54, 57.
2 唐俊易. 服务器托管和服务器租用的区别［J］. 计算机与网络, 2015, 41（1）: 45.
3 陈海锋. 医疗信息化下运营商云服务探析［J］. 电脑知识与技术, 2015, 11（19）: 211 - 213.
4 杨秀峰, 曹晓均, 周毅, 等. 全院信息系统基于公有云托管服务的探索与实践［J］. 中国数字医学, 2016, 11（4）: 90 - 92, 89.
5 张耀祥. 云计算和虚拟化技术［J］. 计算机安全, 2011,（5）: 80 - 82.
6 yien. 服务器托管的双线路接入方法介绍［J］. 计算机与网络, 2013, 39（24）: 37.
7 陈斌. 多线路智能 DNS 系统的设计［J］. 电脑知识与技术, 2014, 10（29）: 6816 - 6817, 6823.
8 朱杰. 智能 DNS 优缺点分析［J］. 信息与电脑（理论版）, 2015,（21）: 94 - 95.