

面向智能健康的人脸表型 - 疾病 - 遗传知识库设计与构建^{*}

何孟桥 谢 雯 张影波 吴蓉蓉 张 珂 詹超英 沈百荣

(四川大学华西医院疾病系统遗传研究院 成都 610212)

〔摘要〕 目的/意义 明确人脸表型 - 疾病 - 基因的复杂关系, 探索三者相互作用机制。方法/过程 以 PubMed 数据库中的科学文献为基础, 综合利用自然语言处理工具和人工过滤方式对现有文献报道的人脸表型 - 疾病 - 基因概念及关系进行知识抽取, 并构建知识库。结果/结论 完成面向智能健康的人脸表型 - 疾病 - 遗传知识库的框架设计与构建, 为探索人脸表型 - 疾病 - 基因相互作用机制和人脸表型在疾病诊断中的应用奠定数据和理论基础。

〔关键词〕 人脸表型; 基因; 知识库; 智能健康

〔中图分类号〕 R - 058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.12.014

Design and Construction of a Face Phenotype - Disease - Genotype Association Knowledge Base for Intelligent Health Application

HE Mengqiao, XIE Wen, ZHANG Yingbo, WU Rongrong, ZHANG Ke, ZHAN Chaoying, SHEN Bairong

Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu 610212, China

〔Abstract〕 **Purpose/Significance** To elucidate the complex relationship among face phenotype - disease - gene, and to explore the underlying mechanisms of their interactions. **Method/Process** Based on scientific literature from the PubMed database, natural language processing tools and manual filtering methods are used to extract the knowledge of the concept and relationship of face phenotype - disease - gene reported in the existing literature, and a knowledge base is constructed. **Result/Conclusion** The study completes the framework design and construction of the knowledge base of the face phenotype - disease - genotype for intelligent health application, which lays a foundation of both data and theory for exploring the interplay between face phenotype - disease - gene, as well as the potential application of face phenotype in disease diagnosis.

〔Keywords〕 human face phenotype; gene; knowledge base; intelligent health

1 引言

人脸表型是由不同面部特征 (包括眼睛、鼻

子、下巴和嘴巴等) 构成的复杂表型体系, 其大小、形状和比例不仅具有显著的遗传性, 而且呈现明显的疾病关联性。在一个克鲁宗综合症的家族内部, 父母和子女的面部特征有很高的重叠度, 都具有下颌前突、浅眼眶、突出的眼球、外斜视等面部表型^[1], 特殊的面部特征在临床疾病诊断中发挥至关重要的作用。

针对人脸面部特征与遗传关系的研究表明, 人类的面部形态特征具有很高的可遗传性, 相关研究

〔修回日期〕 2023 - 06 - 05

〔作者简介〕 何孟桥, 硕士研究生; 通信作者: 沈百荣, 博士, 教授。

〔基金项目〕 国家自然科学基金资助项目 (项目编号: 32270690)。

表明面部特征的遗传力高达 80%^[2]。当遗传决定因素确定时，在一定程度上可以预测一个人的面部特征。患者个体异质性导致疾病临床表型繁多，使疾病的诊断更加困难。Rubinstein - Taybi 综合征是由 CREBBP 基因突变引起的疾病，但在该基因的 30 外显子最后部分和 31 外显子开始部分发生的突变，与一般的 Rubinstein - Taybi 综合征患者表型不同，具有许多非特异性表型^[3]。对于此类复杂疾病，医生难以通过个体的临床表型或基因检测信息给出准确的诊断结果和后期治疗方案。

知识库是用于知识管理的数据库，结合某一领域的特点制订个性化的标准，将非结构化且复杂的知识和信息通过该标准进行结构化，形成知识集合。在生物医学领域，专业知识库的构建通常有两种方法：自然语言处理（natural language processing, NLP）自动构建^[4]和人工数据抽取手动构建^[5-6]。前者由模型进行数据抽取，效率高，但数据准确率低，且数据可用性不高；后者数据抽取质量高，但时间成本很高。因此，构建知识库需要充分利用 NLP 技术，并结合人工审查，以提高数据的可靠性和准确性。

针对上述问题，设计人脸表型 - 疾病 - 基因知识框架，结合 NLP 技术，整合和规范人脸表型 - 疾病 - 基因的相关研究数据，构建人脸表型 - 疾病 - 遗传知识库。该知识库为分析人脸表型 - 疾病 - 基因相互作用关系提供平台基础，为复杂疾病病因的解释提供全面的数据支撑，并将其演化为对应的知识图谱，嵌入人工智能领域，提高相应模型在疾病诊断方面的可解释性。

2 知识库设计与构建

2.1 知识库框架

知识库框架是知识库的基础与核心。在构建知识库初期，需要根据特定的需求，设计相应的知识框架，以便整合和规范文献中的非结构化知识。因为本研究要探索人脸表型 - 疾病 - 基因关系，所以知识库以收集人脸表型 - 疾病 - 基因三元关系为核心，形成知识框架。共定义 4 种类型的实体：参考

文献、人脸表型、基因、疾病。每种实体定义若干属性，如参考文献实体定义了文献标题、作者、期刊以及发表时间 4 种属性，见图 1。

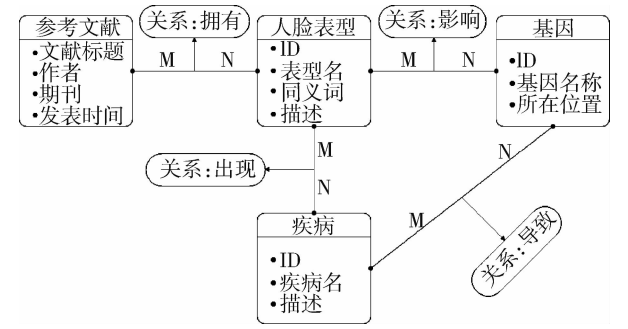


图 1 知识库框架

底层数据管理通过 MySQL 技术实现，使用 SQL 语句创建 4 张表，分别存储 4 种实体数据。4 张表之间存在多对多的关系，因此创建中间表存储关系，形成不同的关系对，见表 1。最终，人脸表型 - 疾病 - 遗传知识库共涉及 8 张表，包括 4 张实体表和 4 张关系表。

表 1 实体关系对

实体 1	关系	实体 2	关系属性
参考文献	拥有	人脸表型	-
基因	影响	人脸表型	分类、证据
疾病	出现	人脸表型	-
基因	导致	疾病	详细信息、分类

2.2 数据收集

2.2.1 数据来源 以 PubMed^[7] 为数据来源。使用高级检索准确定位所需文献。自定义相应检索范式，以人脸表型及其同义词与基因为主要检索内容，辅以其他限制条件，例如：排除综述，规定发表时间区间等。根据人类表型术语集（Human Phenotype Ontology, HPO）^[8] 中人类面部表型本体总结所需人脸表型，最终形成检索式为（face phenotype [tiab]）AND（genotype [tiab]）NOT review [pt]，例如：（high forehead [tiab] OR tall forehead [tiab]）AND（variant * [tiab] or SNP [tiab] or allele * [tiab] or mutat * [tiab] or polymorphism * [tiab]

or haplotype * [tiab] or SNPs [tiab] or genotype [tiab] or gene [tiab]) NOT review [pt] AND (1980 - 01 - 01 [DP] : 2023 - 01 - 01 [DP]) 。

2.2.2 文献处理任务 传统人工方式构建知识库耗时费力，NLP 技术逐渐应用于知识库构建，其中命名实体识别（named entity recognition，NER）、关系抽取（relation extraction，RE）是自动化构建知识库的关键任务。目前在生物医学领域实现 NER 和 RE 的方法可分为基于神经网络的方法、基于多任务学习的方法、基于迁移学习的方法，见表 2。

表 2 NER 和 RE 的方法和工具

类别	方法/工具	任务
基于神经网络	CNN + LSTM ^[9]	NER
	ETM ^[10]	NER
	MCCNN ^[11]	RE
	BERE ^[12]	RE
基于多任务学习	DTranNER ^[13]	NER
	MTL - LBC ^[14]	NER
	Joint AB - LSTM ^[15]	RE
基于迁移学习	BioBERT ^[16]	NER、RE
	HUNER ^[17]	NER

在 NER 方面，生物医学名词一般由多个名词组成，导致医学实体名词长度增加，实体边界识别难度加大^[18]。Lee J 等^[16]提出一种生物医学文本挖掘的预训练生物医学语言表示模型——BioBERT，在生物医学领域的 NER 和 RE 方面表现明显优于其他模型。本研究标注面部表型，例如下颌前突，使用自然语言处理工具包（natural language toolkit，NLTK）对这类表型词及其同义词进行标注和词性标记，作为 BioBERT 模型的训练数据，以此微调 BioBERT 模型，实现对面部表型实体的识别任务。

在 RE 方面，实际应用中既需要模型判断实体间是否存在关系，还需要定义具体的实体间关系。例如，研究^[19]表明 FGF23 基因与下颌前突面部表型有关系，在关系抽取时，只得到一个二分类结果。因此需要人工介入，提取二分类结果为“有关系”文献的具体关系，着重查看实体处于一段话中的情况，如“当 FGF23 基因发生 c.35C > A 突变时，对下颌前突有很强的影响”。如果仍存在差异则进一步分析解决。目前 NLP 还不能完全自动化构建知识库，需要人工校正，对收集的数据二次复查。文献处理任务流程，见图 2。

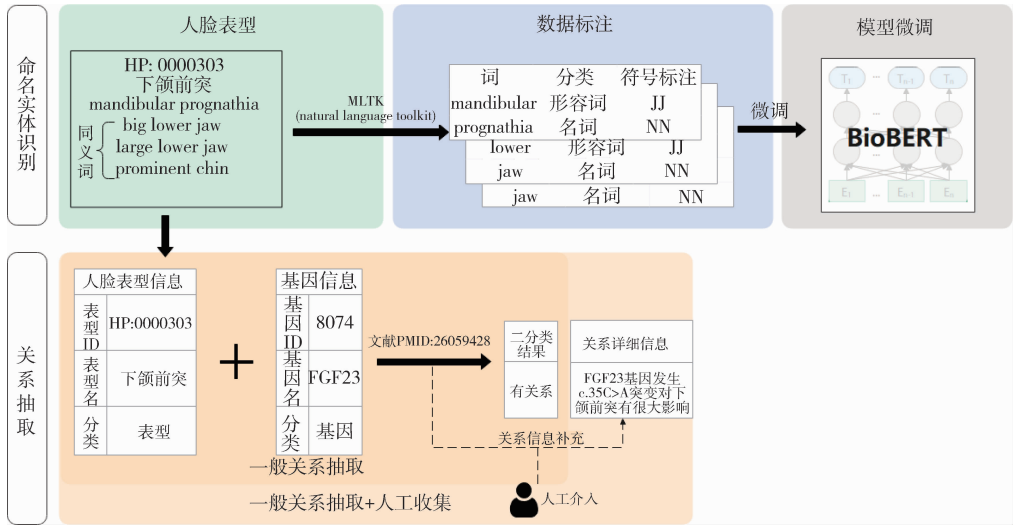


图 2 文献处理任务

2.2.3 半自动化数据检索与收集 人脸表型相关概念多达 2 000 余条，对应的同义词达到 5 000 余条。逐个人工检索这些概念十分耗时且烦琐，为了

提高效率，运用 PubMed 的 Entrez Programming Utilities^[20]中的 Esearch API 进行批量检索，获取相关文献信息。使用 Python 编程语言，编写程序调用 Es-

earch API 接口在 PubMed 数据库中检索人脸表型 - 疾病 - 基因的相关文献。通过文献编号 ID 去重。获取大量相关文献后, 首先, 采用人工阅读方式进行筛选, 排除动物实验和无用数据, 得到 313 篇文献。其次, 为了实现命名实体识别功能, 采用 2022 年 6 月发布的 HPO 版本, 利用 PhenoTagger^[21] 工具提供的代码构建相应表型字典, 并生成训练数据集。再次, 使用该数据集微调 BioBERT 模型, 平均精度为 0.712 50, 平均 *F1* 分数为 0.723 65, 宏平均精度为 0.708 80, 宏平均 *F1* 分数为 0.733 46。最后, 使用 PubTator^[22] 和 PhenoTagger 工具, 将文献

集合作为输入, 得到最终示例数据, 见图 3。数据收集流程, 见图 4。



图 3 示例数据

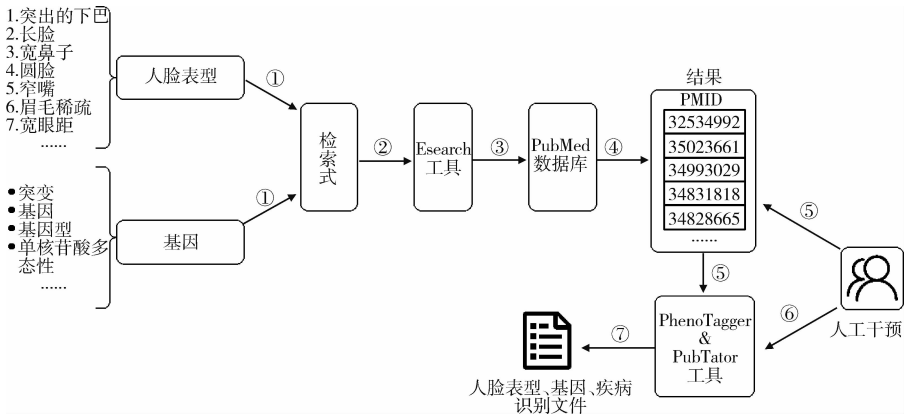


图 4 数据收集流程

获取文献中的人脸表型、基因和疾病实体后, 仍需人工审查, 以提高实体抽取的精确度。同时为了构建知识库, 要定义实体之间的关系。目前关系抽取模型效果不足以满足本研究的个性化需求, 而且关系抽取数据集相对命名实体识别数据集较少, 导致抽取关系的准确率整体低于命名实体识别的准确率。因此, 选择人工方式进行数据收集和整理。最终人脸表型 - 疾病 - 遗传知识库包含人脸表型实体 2 183 个, 基因实体 206 个, 疾病实体 135 个, 参考文献实体 313 个, 相关关系包括 2 221 对“文献 - 人脸表型”关系, 2 773 对“基因 - 人脸表型”关系, 781 对“基因 - 疾病”关系, 1 741 对“人脸表型 - 疾病”关系。

2.3 知识系统搭建

为提供更好的用户交互体验, 在人脸表型与基

因知识库构建基础上进行知识系统搭建。该系统采取前后端分离模式, 主要由前端、后端和数据库部分组成。其中前端部分搭建采用 React 框架作为系统页面的主体框架, 使用 JavaScript 编程语言编写用户交互逻辑, CSS 与 HTML 设计系统页面的展示效果; 后端部分采用 Django 技术框架, 使用 Python 编程语言编写对底层数据的操作逻辑, 同时利用 Django 提供的 Model 类实现后端与数据库的连接, 简化后端对数据库中数据的操作; 数据库部分即知识库部分, 利用 MySQL 技术和 Navicat 工具进行数据存储和管理, 为知识系统提供数据支持。

同时, 针对该知识库特征设计两个主要功能: 数据检索和数据可视化。数据检索功能允许用户查询关于人脸表型、基因以及疾病的详细信息及三者之间的关系, 提供标准化数据供用户下载共享。数据可视化功能利用 Echarts 或 D3.js 等插件实现。通过旭日图、

桑基图、关系图等图表展示数据，使数据关系更加直观，用户交互体验更好。例如，用户可以通过桑基图观察基因与人脸表型的作用关系，通过旭日图了解对面部表型有影响的基因在染色体上的具体位置。知识系统整体技术框架和主要功能，见图 5。

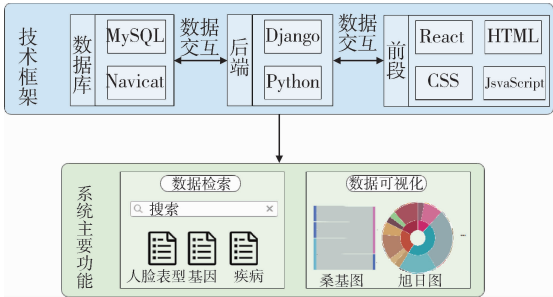


图 5 知识系统技术框架和主要功能

3 人脸表型 - 疾病 - 遗传知识库的应用潜力

构建该人脸表型 - 疾病 - 遗传知识库，实际上

是构建了一个人脸表型、疾病与基因的系统遗传网络，通过它能更好地分析面部表型、基因以及疾病之间的复杂关系，提供一定的数据参考，研究不同基因与面部表型特征之间的关系，探索疾病的发病机制。数据收集过程也涉及许多罕见病，一些罕见病患者的面部表型特殊，具有明显特点，该知识库有望通过面部表型数据，对这些患者辅助诊断，为复杂疾病的病因提供相应的证据支持。在该知识库中以 KBC 综合征为例，ANKRD11 基因的突变^[23-25]造成 KBC 综合征，同时患者显示出大牙齿、长人中、眼距过远、上翘的鼻子、宽头、上睑下垂等面部表型。

随着人脸表型 - 疾病 - 遗传知识库不断更新，积累更多相关知识，可以将其演化为相应的知识图谱，预测未知的人脸表型、疾病以及基因关系，推理未知疾病的遗传机制和可能产生的表型，也可以将其应用于聊天机器人，通过多轮对话，训练人脸表型 - 疾病 - 基因知识图谱模型，使其给出的答案更加准确，见图 8。

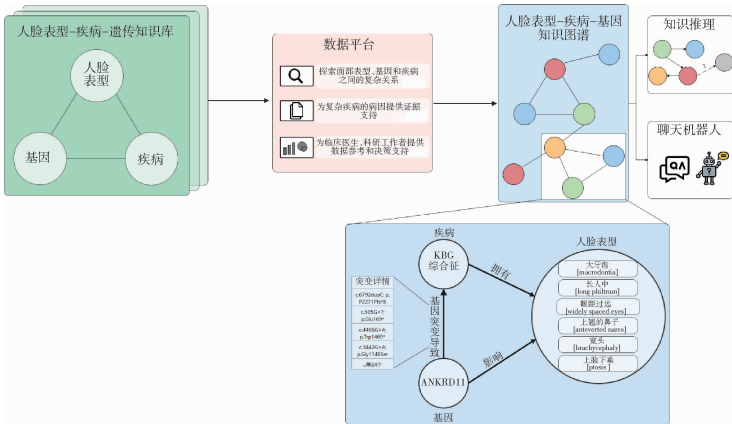


图 8 知识库应用

4 结语

随着各领域的知识不断积累，知识库已经成为管理和整合知识不可或缺的工具。构建一个医学知识库是具有挑战性的，因为医学知识本身十分复杂，而且信息抽取技术尚不成熟，难以达到预期效果。本研究采用人工方式和 NLP 工具，收集和整理文献中关于人脸表型、基因和疾病的知识，并利用

计算机和数据库技术构建相应知识库。该知识库形成了一个系统的知识网络，可用于提供个性化的知识检索和标准化数据集，为临床医生和科研工作者提供数据参考和决策支持。然而，信息抽取仍要辅以人工进行二次检查，知识系统也要在后期进行功能完善和增加。

在未来的研究中，人脸表型 - 疾病 - 遗传知识库的知识将不断扩充，逐渐演化为医学本体和知识图谱，将其嵌入可解释人工智能板块，能推动人工

智能可解释性的发展。该知识库有望在法医鉴定、临床辅助诊断等领域得到应用。

参考文献

- LI Z L, CHEN X, ZHUANG W J, et al. FGFR2 mutation in a Chinese family with unusual Crouzon syndrome [J]. *International journal of ophthalmology*, 2016, 9 (10): 1403 – 1408.
- JOHANNSDOTTIR B, THORARINSSON F, THORDARSON A, et al. Heritability of craniofacial characteristics between parents and offspring estimated from lateral cephalograms [J]. *American journal of orthodontics dentofacial orthopedics*, 2005, 127 (2): 200 – 207.
- MENKE L A, VAN BELZEN M J, ALDERS M, et al. CREBBP mutations in individuals without Rubinstein – Taybi syndrome phenotype [J]. *American journal of medical genetics – part A*, 2016, 170 (10): 2681 – 2693.
- WANG Y, WANG Y, WANG S, et al. GIDB: a knowledge database for the automated curation and multidimensional analysis of molecular signatures in gastrointestinal cancer [EB/OL]. [2023 – 06 – 01]. <https://doi.org/10.1093/database/baz051>.
- BI C, ZHOU S, LIU X, et al. NDDRF: a risk factor knowledgebase for personalized prevention of neurodegenerative diseases [EB/OL]. [2023 – 06 – 01]. <https://doi.org/10.1016/j.jare.2021.06.015>.
- ZHAN C, SHI M, WU R, et al. MIRKB: a myocardial infarction risk knowledge base [EB/OL]. [2023 – 06 – 01]. <https://doi.org/10.1093/database/baz125>.
- 吕霖. PubMed 及其衍生数据库在医药领域的应用 [J]. *中国发明与专利*, 2014 (8): 107 – 113.
- ROBINSON P N, KOHLER S, BAUER S, et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease [J]. *American journal of human genetics*, 2008, 83 (5): 610 – 615.
- CHO M, HA J, PARK C, et al. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition [EB/OL]. [2023 – 06 – 01]. <https://doi.org/10.1016/j.jbi.2020.103381>.
- DAI X, KARIMI S, HACHEY B, et al. An effective transition – based model for discontinuous NER [C]. *online: Association for Computational Linguistics*, 2020.
- QUAN C, HUA L, SUN X, et al. Multichannel convolutional neural network for biological relation extraction [J]. *BioMed research international*, 2016 (25): 1850404.
- HONG L, LIN J, LI S, et al. A novel machine learning framework for automated biomedical relation extraction from large – scale literature repositories [J]. *Nature machine intelligence*, 2020, 2 (6): 347 – 355.
- HONG S K, LEE J G. DTranNER: biomedical named entity recognition with deep learning – based label – label transition model [J]. *BMC bioinformatics*, 2020, 21 (1): 53.
- ZUO M, ZHANG Y. Dataset – aware multi – task learning approaches for biomedical named entity recognition [J]. *Bioinformatics*, 2020, 36 (15): 4331 – 4338.
- SAHU S K, ANAND A. Drug – drug interaction extraction from biomedical texts using long short – term memory network [EB/OL]. [2023 – 06 – 01]. <https://doi.org/10.1016/j.jbi.2018.08.005>.
- LEE J, YOON W, KIM S, et al. BioBERT: a pre – trained biomedical language representation model for biomedical text mining [J]. *Bioinformatics*, 2020, 36 (4): 1234 – 1240.
- WEBER L, MÜNCHMEYER J, ROCKTÄSCHEL T, et al. HUNER: improving biomedical NER with pretraining [J]. *Bioinformatics*, 2019, 36 (1): 295 – 302.
- 谭玲, 鄂海红, 匡泽民, 等. 医学知识图谱构建关键技术及研究进展 [J]. *大数据*, 2021, 7 (4): 80 – 104.
- CHEN F, LI Q, GU M, et al. Identification of a mutation in FGF23 involved in mandibular prognathism [J]. *Scientific reports*, 2015, 5 (1): 11250.
- SAYERS E W, BOLTON E E, BRISTER J R, et al. Database resources of the national center for biotechnology information [J]. *Nucleic acids research*, 2022, 50 (D1): D20 – D26.
- LUO L, YAN S, LAI P T, et al. PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology [J]. *Bioinformatics*, 2021, 13 (37): 1884 – 1890.
- WEI C H, ALLOT A, LEAMAN R, et al. PubTator central: automated concept annotation for biomedical full text articles [J]. *Nucleic acids research*, 2019, 47 (W1): W587 – W593.
- LI Q, SUN C, YANG L, et al. Comprehensive analysis of clinical spectrum and genotype associations in Chinese and literature reported KBG syndrome [J]. *Translational pediatrics*, 2021, 10 (4): 834 – 842.
- OCKELOEN C W, WILLEMSSEN M H, DE MUNNIK S, et al. Further delineation of the KBG syndrome phenotype caused by ANKRD11 aberrations [J]. *European journal of human genetics*, 2015, 23 (9): 1176 – 1185.
- MURRAY N, BURGESS B, HAY R, et al. KBG syndrome: an Australian experience [J]. *American journal of medical genetics – part A*, 2017, 173 (7): 1866 – 18773.