# pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization and comparison

Zhi-Hui Luo[1], Meng-Wei Shi[1], Zhuang Yang[1], Hong-Yu Zhang[2]*, Zhen-Xia Chen[1]*

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, PR China

[2]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P. R. China.

*To whom correspondence should be addressed.

**Contact:** zhen-xia.chen@mail.hzau.edu.cn or zhy630@mail.hzau.edu.cn

**Running head**: MeSH toolkit pyMeSHSim

**Key words**: MeSH, UMLS, named entity recognition, semantic similarity, supplementary concept records, disease

**ABSTRACT**

**Motivation:** Increasing disease causal genes have been identified through different methods, while there are still no uniform biomedical named entity (bio-NE) annotations of the disease phenotypes. Furthermore, semantic similarity comparison between two bio-NE annotations, like disease descriptions, has become important for data integration or system genetics analysis.

**Methods:** The package pyMeSHSim realizes bio-NEs recognition using MetaMap, which produces Unified Medical Language System (UMLS) concepts in natural language process. To map the UMLS concepts to MeSH, pyMeSHSim embedded a house made dataset containing the Medical Subject Headings (MeSH) main headings (MHs), supplementary concept records (SCRs) and relations between them. Based on the dataset, pyMeSHSim implemented four information content (IC) based algorithms and one graph-based algorithm to measure the semantic similarity between two MeSH terms.

**Results:** To evaluate its performance, we used pyMeSHSim to parse OMIM and GWAS phenotypes. The inclusion of SCRs and the curation strategy of non-MeSH-synonymous UMLS concepts used by pyMeSHSim improved the performance of pyMeSHSim in the recognition of OMIM phenotypes. In the curation of GWAS phenotypes, pyMeSHSim and previous manual work recognized the same MeSH terms from 276/461 GWAS phenotypes, and the correlation between their semantic similarity calculated by pyMeSHSim and another semantic analysis tool *meshes* was as high as 0.53-0.97.

**Conclusion:** With the embedded dataset including both MeSH MHs and SCRs, the integrative MeSH tool pyMeSHSim realized the disease recognition, normalization and comparison in biomedical text-mining.

**Availability:** Package's source code and test datasets are available under the GPLv3 license at https://github.com/luozhhub/pyMeSHSim

## INTRODUCTION

Biomedical Named entity (bio-NE) recognition, normalization and comparison are fundamental and important tasks for extracting and utilizing valuable biomedical knowledge from textual data (1-4). They are realized by identifying key entities in unstructured text, mapping identified entities to a controlled vocabulary, and measuring the semantic similarity between the controlled vocabulary terms (5).

Medical Subject Headings (MeSH) is a controlled vocabulary that can be used in bio-NE recognition, normalization and comparison (6). It consists of three main record types including descriptor records, qualifier records and supplementary concept records (SCRs). MeSH is curated by the National Library of Medicine (NLM) and serves as the index system in PubMed/MEDLINE and other NLM databases. NLM has used Medical Text Indexer (MTI) to provide indexing recommendations based on MeSH in the bio-NE recognition for literatures since 2002 (7). As precise literature annotations, MeSH has become more and more popular in the normalization of bio-NEs, including disease names, in medical and genetic public databases (8, 9). Also, the structure of MeSH as a directed acyclic graph like Gene Ontology (10) and Disease Ontology (11) enables the comparison of semantic similarity between two MeSH terms in the graph.

Several MeSH tools have been developed to realize bio-NE recognition, normalization or comparison. As a MeSH tool for bio-NE recognition and normalization, NLM MeSH has provided a browser online (https://meshb.nlm.nih.gov/search) to parse MeSH terms from the input phrases. However, it is neither tolerant to even subtle difference of input phrases from MeSH terms, nor applicable to batch processing. As MeSH tools for bio-NE comparison, *meshes* (12) and *meshSim* (13) have been developed recently to measure MeSH semantic similarity using the R dataset MeSH.Hsa.eg.db (3) as data framework. However, the lack of SCRs embedded in its MeSH dataset limits the use of both bio-NE comparison tools for rare diseases such as “alzheimer’s disease 7” and “Bardet-Biedl syndrome 11”. Furthermore, there is still a lack of an integrated one-stop MeSH toolkit to realize bio-NE recognition, normalization and comparison.

To solve above problems, an integrative python package pyMeSHSim was developed to realize bio-NE recognition, normalization and comparison. It can directly curate MeSH terms from free biomedical descriptions and measure the semantic similarity between the MeSH term pairs. Additionally, to enable batch processing and the application of pyMeSHSim to both common diseases and rare diseases, a lightweight and comprehensive MeSH dataset, was generated and embedded as the data framework of pyMeSHSim.

## METHODS

### Dataset construction

A comprehensive MeSH dataset is fundamental to MeSH tools, while the MeSH dataset used

by most popular MeSH tools contains only MeSH Main Headings (MHs), a component of MeSH descriptor records, but not SCRs. To construct a comprehensive MeSH dataset, we exacted MeSH information, including MHs, SCRs, and their relations, from Unified Medical Language System (UMLS, 2018AA version), which is a large biomedical thesaurus integrated nearly 200 vocabularies including MeSH (14).

The multiple-to-one relationship between MeSH-synonymous UMLS concepts and MeSH MHs was curated from the UMLS table MRSAT. For example, the MeSH MH “Alzheimer Disease” (D000544) includes seven MeSH concepts, each of which corresponds to several MeSH entry terms and a UMLS concept (Supplementary table 1). In our dataset, we included the MeSH MHs and related UMLS concepts, while excluded the MeSH concept and MeSH entry term information. Moreover, we curated the most useful “parent” and “child” relationship between MeSH MHs from the UMLS table MRREL.

The one-to-one relationship between MeSH-synonymous UMLS concepts and SCRs was curated from the UMLS table MRSAT. We included the SCRs and its corresponding UMLS concepts, as well as the “narrower” and “broader” relationship between SCRs and MeSH MHs curated from the UMLS table MRREL, in our dataset.

The qualifier records and other MeSH descriptor records except MeSH MHs were not included in our dataset. We used “MeSH term” in the following to refer MeSH MH or SCR.

**Bio-NE recognition and normalization**

The bio-NE recognition were realized by MetaMap (15), a widely used biomedical natural language processing software that recognizes UMLS concepts from free texts. Although machine learning methods might have better performance than MetaMap in recommending MeSH MHs to MEDLINE citations, their use were constrained by the requirement of large amount of training data to establish the model, and the potential imbalance of the training data (16). Meanwhile, the bio-NEs, such as disease phenotypes from GWASdb (17), OMIM (18), and GAD (19), and drug indications from public databases DrugBank (20) and TTD (21), might have to be curated from free texts without large amount of training data, which is the expertise of MetaMap. The UMLS concepts curated by MetaMap were then converted to MeSH terms based on our dataset. MeSH-synonymous UMLS concepts were mapped to MHs or SCRs, and non-MeSH-synonymous UMLS concepts were curated as free texts and mapped to MHs or SCRs.

**Bio-NE comparison**

We compared the bio-NEs based on the similarity between their corresponding MeSH terms. The semantic similarity is usually calculated by graph-based or information content (IC)-based measures. The graph-based method measures the path distance between two MeSH terms on the MeSH hierarchical structure, while the IC-based method depends on the specificity and informativeness of MeSH terms (22).

We retrieved the number of publications indexed by MeSH terms using the NCBI E-Utility (23), and calculated the IC vaules as below.

$$D(d) = \{Descendants\ of\ d\} \quad (1)$$

$$P(d) = \frac{freq(D(d))}{N} \quad (2)$$

$$IC(d) = -log\ (P(d)) \quad (3)$$

Where *D(d)* is the descendent terms of MeSH term *d*; *freq(x)* is the number of publications indexed by term *x*; *N* is the total number of publications indexed by MeSH; and *IC(d)* is the IC value of term *d*.

We implemented the following four IC-based algorithms:

$$Sim_{res}(d_1,\ d_2) = IC(MICA\{d_1,\ d_2\}) \quad (4)$$

$$Sim_{lin}(d_1,\ d_2) = \frac{2 \times IC(MICA\{d_1, d_2\})}{IC(d_1) + IC(d_2)} \quad (5)$$

$$Sim_{JC}(d_1,\ d_2) = 1 - min\big(1, IC(d_1) + IC(d_2) - 2 \times IC(MICA\{d_1,\ d_2\})\big) \quad (6)$$

$$Sim_{rel}(d_1,\ d_2) = Sim_{lin}(d_1,\ d_2) \times \left(1 - 10^{-IC(MICA\{d_1, d_2\})}\right) \quad (7)$$

Where $d_1$ and $d_2$ are MeSH terms; $Sim_{lin}$, $Sim_{res}$, $Sim_{rel}$, $Sim_{JC}$ corresponds to Lin's (24), Resnik's (25), Schlicker's (26), and Jiang and Conrath's (27) algorithms, respectively; MICA (the most informative common ancestor) is the ancestor of the selected two MeSH terms with the maximal IC value among all ancestors. We assigned MICA between MeSH terms from different categories, denoted by the first character of the tree number of MeSH terms, to 0. For example, MICA between the MeSH terms “Tauopathies” (tree number: “C10.574.945”) and “Schizophrenia” (tree number: “F03.700.750”) is 0 because they belong to different categories (“C” for diseases vs “F” for psychiatry and psychology).

We also implemented the graph based Wang’s (28) algorithm as below.

$$A(d) = \{Ancestor\ of\ d\} \quad (8)$$

$$S_d(a) = max\{\omega^{n_a}\}, a\ \in A(d) \quad (9)$$

$$SV_d = \sum_{t \in A(d)} S_d(t) \quad (10)$$

$$Sim_{Wang}(d_1,\ d_2) = \frac{\sum_{t \in A(d_1) \cap A(d_2)} \left(S_{d_1}(t) + S_{d_2}(t)\right)}{SV_{d_1} + SV_{d_2}} \quad (11)$$

Where *d* is a MeSH term; *A(d)* is the ancestors, deduced from tree numbers, of *d*; $\omega$, which is set to 0.7 in our work, is the weight used to measure the relation between two terms; $n_a$ is the number of edges between *d* to *a*; $S_d(a)$ is the semantic contribution of *a* to *d*; $SV_d$ is the total semantic contributions of all ancestors to *d*; $Sim_{Wang}(d_1, d_2)$ is Wang's algorithm score between MeSH terms $d_1$ and $d_2$.

Noteworthily, both IC-based and graph-based methods depend on the tree numbers, which may be not unique, of the MeSH terms, and thus lead to multiple similarity values. We retained only the maximal similarity value between two MeSH terms.

## IMPLEMENTATION

The pyMeSHSim consists of three subpackages (1) the *metamapWrap* subpackage that recognizes bio-NEs from the text, and converts bio-NEs into MeSH terms in bio-NE recognition (2) the *data* subpackage translates UMLS concepts to MeSH terms, by the embedded MeSH dataset in bio-NE normalization, and (3) the *Sim* subpackage that measures the distance between MeSH terms in the bio-NE comparison (Figure 1).

### The metamapWrap subpackage

The bio-NE recognition and normalization of pyMeSHSim were realized by the *metamapWrap* subpackage, which was a wrapper for MetaMap (15). The subpackage *metamapWrap* curated MeSH-synonymous UMLS concepts from free texts, including non-MeSH-synonymous UMLS concepts, and then converted UMLS concepts to corresponding MeSH terms via the data subpackage. We set parameters "-N –J list_semantic_type –R MSH –I –z –conj –Q 4 –silent --sldi", where list_semantic_type was the list of disease-related semantic types "inpo,dsyn,phpranab,orgf,clna,hlca,genf,orga,neop,emod,inbe,lbtr,anst,npop,celc,cell,bpoc,acty,mobd,celf,evnt,sosy,patf,tisu,moft,fndg,bdsu,ortf,menp,acab,comd,sbst,cgab", as default of pyMeSHSim (see manuals at http://pymeshsim.systemsgenetics.cn/). Users can tune a series of parameters to get more accurate results.

### The data subpackage

The MeSH dataset was embedded in the *data* subpackage in bcolz format with a corresponding data interface (Supplementary table 2). It included five tables: (1) table "MainHeadingDetailData" stored all the MH information, including MeSH unique id, tree code, prefer name, category, term semantic type, IC frequency, and UMLS id. The semantic type was curated from the UMLS table MRSTY, and each UMLS concept was characterized by at least one of the 133 semantic types (29). (2) table "supplementMainHeading" stored all the UMLS concepts related to MHs; (3) table "RNDetailData" stored the basic information of SCRs; (4) table "RNandRBRel" stored the narrower-and-broader relationship between SCRs and MHs; and (5) table "ParentChildRel" stored the fundamental tree structure. The five tables enabled the conversion of UMLS concepts to MeSH terms, and the measurement of the semantic similarity between MeSH terms.

### The Sim subpackage

The bio-NE comparison of pyMeSHSim was realized by the *Sim* subpackage via measuring the distance between MeSH terms. Each narrower record from the SCRs was converted to one or more broader terms from MHs before the measurement. As the tool *meshes*, pyMeSHSim offered five representative methods of semantic similarity measures, including four information content (IC) based (Lin's, Resnik's, Schlicker's, and Jiang and Conrath's) and one graph-based (Wang's) algorithms.

## PERFORMANCE

### Evaluation with OMIM phenotypes

To test whether the involvement of SCRs and our curation strategy of non-MeSH-synonymous UMLS concepts would contribute to improved performance of pyMeSHSim in bio-NE recognition, we compared the genes annotated with MeSH MHs and SCRs from OMIM (18) phenotype-gene pairs. The OMIM phenotype-gene pairs were collected from the database disease-connect (30), which used MetaMap to process the disease phenotypes into MeSH-synonymous and non-MeSH-synonymous UMLS concepts. Using pyMeSHSim, MeSH-synonymous UMLS concepts were mapped to MHs and SCRs, which were further mapped to their “broader” MHs, and non-MeSH-synonymous UMLS concepts were curated as free texts and mapped to MHs. We classified OMIM phenotypes into MH, SCR, and non-MeSH groups based on the source of their corresponding UMLS concepts, and compared genes corresponded to the same MHs from all the three groups (Figure 2). The genes without Entrez IDs, which were required for the following disease enrichment analysis, and the MHs with less than 10 genes in two or three groups were excluded from subsequently analysis. After the filtering, 36 MHs and 1498 MH-gene pairs (Supplementary table 3) were remained, including 761 MH-gene pairs from MH group, 552 from SCR group and 215 from non-MeSH group. About 87.5% MH-gene pairs in SCR group also presented in MH group, demonstrating high overlap of genetic features between subtype diseases and its corresponding MH diseases, and validating the significance of SCRs in disease curation (Figure 3). Additionally, 59.5% and 10.6% MH-gene pairs in non-MeSH group presented in MH and SCR group, respectively, indicating the effectiveness of our curation strategy of non-MeSH-synonymous UMLS concepts.

To further validate the reasonability of SCR involvement and our curation strategy of non-MeSH-synonymous UMLS concepts, we explored whether the additional MH-gene pairs deduced from SCRs and non-MeSH-synonymous UMLS concepts were enriched in the MH diseases. We remained the seven MHs with at least five MH-gene pairs outside the MH group, and tested their enriched diseases using the UMLS-based disease enrichment analysis tool DOSE (31). For each of the seven MHs, the addition of genes from SCR and non-MeSH groups led to more significant enrichment in the disease mapped to the MH by pyMeSHSim (Table 1). Especially, the addition of 50 genes from SCR and non-MeSH groups to the 14 genes from MH group, of the MeSH MH “Osteochondrodysplasias (D010009)” led to the

higher pvalue (6.57E-35 vs 8.87E-19) of enrichment in the disease “Osteochondrodysplasias” (Table 1). The more significant enrichment indicated the contribution of SCRs in our dataset and our curation strategy of non-MeSH-synonymous UMLS concepts to the improved performance of pyMeSHSim in bio-NE recognition and normalization.

**Evaluation with GWAS phenotypes**

We compared the performance of pyMeSHSim in the recognition and normalization of bio-NEs from GWAS phenotypes with the manual results of Nelson’s group using MeSH browser (4), and compared their semantic similarity calculated by pyMeSHSim and *meshes* (12).

We firstly used pyMeSHSim to parse the 461 GWAS phenotypes collected by Nelson’s group to MeSH terms. PyMeSHSim successfully curated MeSH terms from 443 (96%) GWAS phenotypes. The curated MeSH terms from 310 (67%) GWAS phenotypes had Lin score > 0.75 with the manual work, demonstrating high performance of pyMeSHSim in bio-NE recognition and normalization (Supplementary table 4).

We then calculated the overall semantic similarity between the curated MeSH terms using two methods by pyMeSHSim and the latest semantic analysis tool *meshes* (Supplementary table 4). The similarity calculated using both packages was 1 between the same terms, and 0 between terms in different categories. For the comparison between the different term pairs of 53 GWAS phenotypes in the same category with Nelson’s group, pyMeSHSim succeeded in calculating the similarities between the term pairs of all the phenotypes, while *meshes*, which only compared MH-MH pairs, failed for SCR-MH pairs of 15 phenotypes (Supplementary table 4). Among the 15 SCRs parsed by pyMeSHSim, 13 were mapped to the same MHs as parsed by Nelson’s group. The similarity correlation of the remaining 38 term pairs between pyMeSHSim and *meshes* was 0.53 (Res’)-0.97 (Lin’s and Rel’s) (Table 2), demonstrating similar, if not better, performance of pyMeSHSim compared with *meshes* in bio-NE comparison.

Comparing the different terms of the 114 GWAS phenotypes curated by pyMeSHSim and the Nelson’s group, we found that the latter preferred mapping the phenotypes to disease category (C). For example, “Color, Eye” was returned by MeSH browser with GWAS phenotype “Eye color” as the input. However, “Eye color” was curated as “color vision defects” by Nelson’s group, while as “Color, Eye” by pyMeSHSim. Similarly, GWAS phenotypes “Hair color” and “Serum urate” were curated as “hair diseases” and “urinary calculi” by Nelson’s group, while as “Color, Hair” and “Acid, Uric” by pyMeSHSim (Supplementary table 4). Therefore, at least a part of the differences between the manual work and pyMeSHSim were from the human bias in the manual work.

In summary, we used OMIM phenotypes and GWAS phenotypes to evaluate pyMeSHSim performance in bio-NE recognition, normalization and comparison. Both evaluations illustrate its high effectiveness and precision in bio-NEs recognition and normalization. GWAS analysis also demonstrated its power in bio-NE comparison. Moreover, OMIM analysis validated its

effectiveness in the integration of genetic information.

## DISCUSSION

We developed pyMeSHSim, an integrative, lightweight and data-rich python package for biomedical text mining. To the best of our knowledge, this is the first universal MeSH toolkit for integrated bio-NE recognition, normalization and comparison analysis. PyMeSHSim is expected to be widely used as a powerful tool in bioinformatics, computational biology and biomedical research.

Considering that MeSH is one of the most widely used biomedical vocabulary, pyMeSHSim will further contribute to data integration. Also, the inclusion of SCRs in the implemented dataset enabled pyMeSHSim to handle rare diseases in public databases like OMIM and Orphanet (www.orpha.net). However, whether general concepts like MHs or more specific concepts like SCRs are preferable will depend on the end use. Users should be cautious to select the right terms accordingly in using pyMeSHSim.

## AUTHOR CONTRIBUTIONS

Z.C. and H.Z. supervised the work. H.Z. and Z.L. conceived the idea. Z.L. and M.S. developed the software. Z.L. and Z.Y. performed the research. Z.L. and Z.C. wrote the manuscript.

## FUNDING

This work was supported by Huazhong Agricultural University Scientific & Technological Self-innovation Foundation [2016RC011]; and the Fundamental Research Funds for the Central Universities [2662018PY021, 2662017PY115].

## REFERENCES


1 Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Science Translational Medicine. 2014;**6**(252):252ra123.

2 Wang H, Gu Q, Wei J, Cao Z, Liu Q. Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. Clinical Pharmacology & Therapeutics. 2015;**97**(5):451.

3 Tsuyuzaki K, Morota G, Ishii M, Nakazato T, Miyazaki S, Nikaido I. MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. Bmc Bioinformatics. 2015;**16**(1):45.

4 Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nature Genetics. 2015;**47**(8):856-60.

5 Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. Journal of biomedical informatics. 2015;**57**:28-37.

6 Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000;**88**(3):265.

7 Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Medinfo. 2004;**89**.

8 Cui T, Zhang L, Huang Y, et al. MNDR v2.0: an updated resource of ncRNA–disease associations in mammals. Nucleic Acids Research. 2018;**46**(Database issue):D371-D4.

9 Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017 Jan 4;**45**(D1):D833-D9.

10 Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;**32**(suppl_1):D258-D61.

11 Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012 Jan;**40**(Database issue):D940-6.

12 Yu G. Using meshes for MeSH term enrichment and semantic analyses. Bioinformatics. 2018;**1**:2.

13 Zhou J, Shui Y, Peng S, Li X, Mamitsuka H, Zhu S. MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents. Journal of bioinformatics and computational biology. 2015;**13**(06):1542002.

14 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;**32**(suppl_1):D267-D70.

15 Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;**17**(3):229-36.

16 Yepes AJJ, Mork JG, Demner-Fushman D, Aronson AR. Comparison and combination of several MeSH indexing approaches. AMIA annual symposium proceedings; 2013: American Medical Informatics Association; 2013. p. 709.

17 Li MJ, Wang P, Liu X, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic acids research. 2011;**40**(D1):D1047-D54.

18 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research. 2014;**43**(D1):D789-D98.

19 Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nature

genetics. 2004;**36**(5):431.

20 Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research. 2006;**34**(suppl_1):D668-D72.

21 Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic acids research. 2002;**30**(1):412-5.

22 McInnes BT, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. AMIA Annual Symposium Proceedings; 2009: American Medical Informatics Association; 2009. p. 431.

23 Sayers E. The E-utilities in-depth: parameters, syntax and more. Entrez Programming Utilities Help [Internet]. 2009.

24 Lin D. An information-theoretic definition of similarity. Icml; 1998: Citeseer; 1998. p. 296-304.

25 Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007. 1995.

26 Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics. 2006;**7**(1):302.

27 Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008. 1997.

28 Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;**23**(10):1274-81.

29 McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Studies in health technology and informatics. 2001;**84**(0 1):216.

30 Liu C-C, Tseng Y-T, Li W, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic acids research. 2014;**42**(W1):W137-W46.

31 Yu G, Wang L-G, Yan G-R, He Q-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2014;**31**(4):608-9.

## TABLES

**Table 1. Disease enrichment analysis of the genes assigned to the MHs before and after addition of MH-gene pairs from SCR and non-MeSH groups.**

| OMIM diseases[1] | | MH-gene pairs (MH group / all)[2] | Enriched UMLS diseases with DOSE | | | *P* value (MH group / all)[4] |
|---|---|---|---|---|---|---|
| MH ID | MH description | | UMLS ID | UMLS description | MH ID[3] | |
| D057130 | Leber Congenital Amaurosis | 17/22 | C0339527 | Leber Congenital Amaurosis | D057130 | 3.43E-33 / 1.45E-42 |
| D020754 | Spinocerebellar Ataxias | 23/28 | C0087012 | Ataxia, Spinocerebellar | D020754 | 1.93E-30 / 2.84E-38 |
| D052177 | Kidney Diseases, Cystic | 19/25 | C1691228 | Cystic Kidney Diseases | D052177 | 8.05E-19 / 2.37E-20 |
| D010009 | Osteochondrodysplasias | 14/64 | C0029422 | Osteochondrodysplasias | D010009 | 8.87E-19 / 6.57E-35 |
| D002925 | Ciliary Motility Disorders | 26/31 | C0008780 | Ciliary Motility Disorders | D002925 | 1.60E-23 / 3.90E-33 |
| D015419 | Spastic Paraplegia, Hereditary | 28/36 | C0037773 | Spastic Paraplegia, Hereditary | D015419 | 1.22E-37 / 2.71E-45 |
| D007938 | Leukemia | 18/51 | C0085669 | Acute leukemia | D007938 | 3.26E-10 / 6.63E-26 |

Notes:

[1] The OMIM diseases were collected from the database disease-connect (34) with at least five MH-gene pairs outside the MH group.

[2] (Number of MH-gene pairs in MH group) / (number of MH-gene pairs in all the three groups including MH, SCR and non-MeSH group).

[3] The MH ID was mapped from the UMLS ID by pyMeSHSim.

[4] (The enrichment P value of genes in MH group) / (The enrichment P value of genes in all the three groups)

**Table 2. Correlation of calculated semantic similarities between pyMeSHSim and *meshes*.**

| **Method** | **Lin's** | **Res'** | **Jiang's** | **Rel's** | **Wang's** |
|---|---|---|---|---|---|
| **Correlation coefficient** | 0.97 | 0.53 | 0.96 | 0.97 | 0.84 |
| ***P* value** | <2.2e-16 | 1.8e-03 | <2.2e-16 | <2.2e-16 | 1.7e-11 |

**FIGURES**

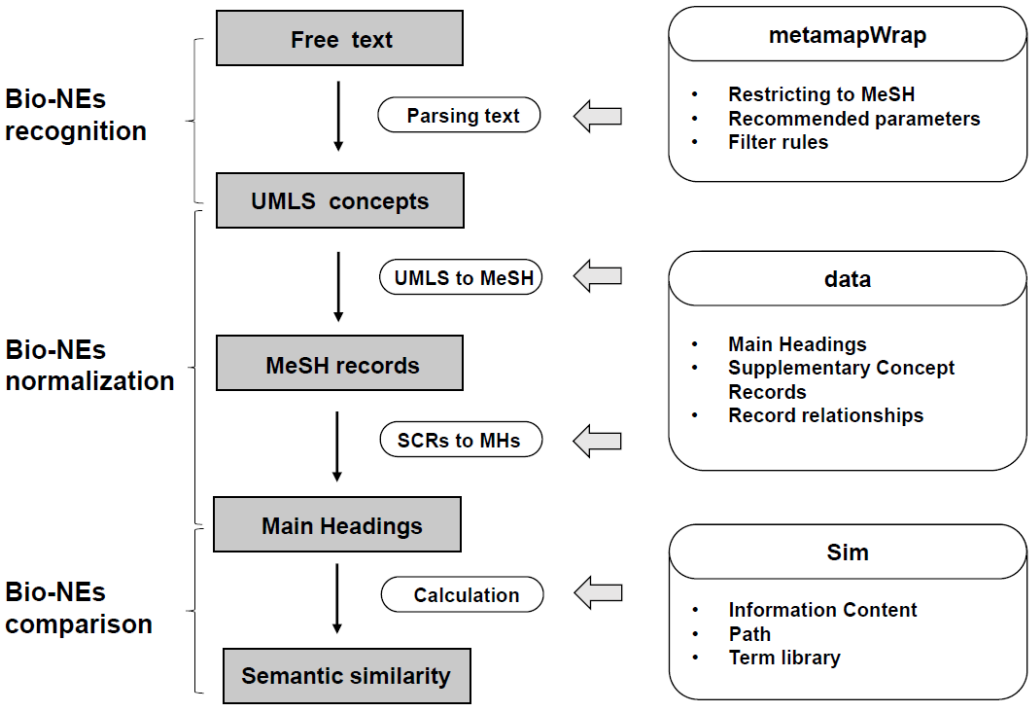


**Figure 1. The components and workflow of pyMeSHSim**. pyMeSHSim consists of three subpackages, including *metamapWrap*, *data* and *Sim*. In bio-NE recognition, *metamapWarp* curates the UMLS concepts from free text. In bio-NE normalization, *data* translates UMLS concepts to MeSH terms, and maps SCRs to MHs using selected records and relationships between records in MeSH. In bio-NEs comparison, *Sim* uses IC-based and graph-based methods to measure semantic similarity between two bio-NEs.

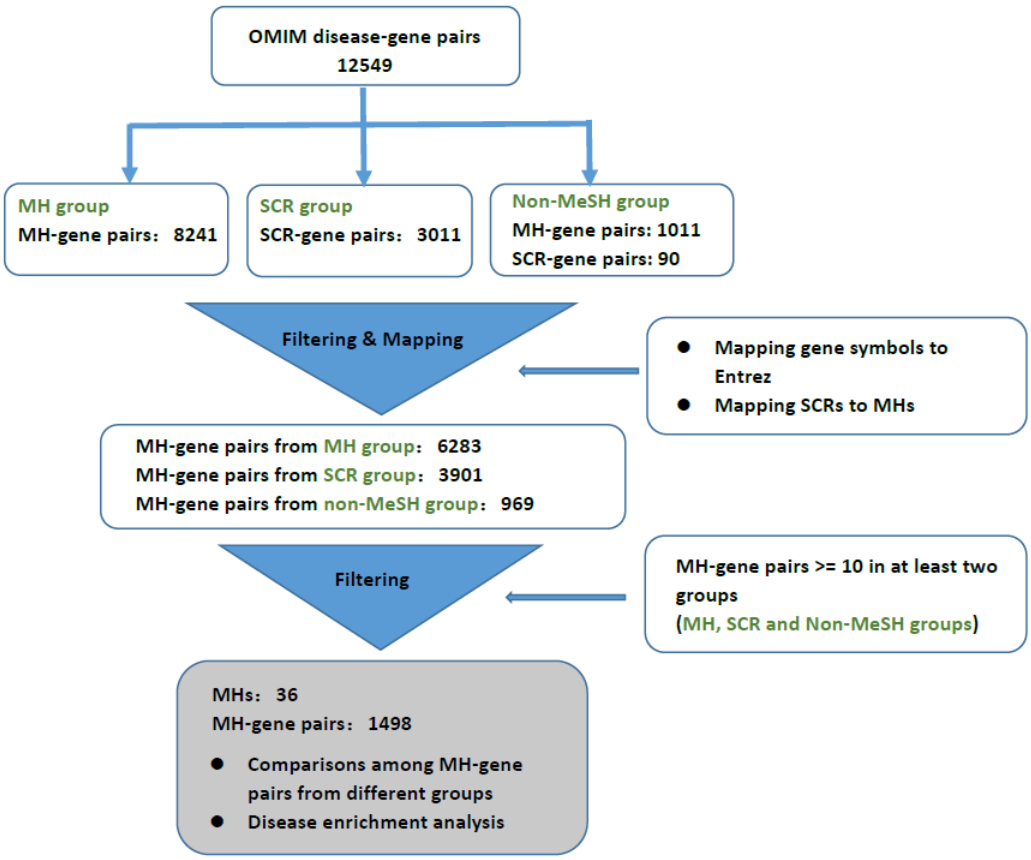


**Figure 2**. **OMIM UMLS diseases processing pipeline.** MeSH-synonymous UMLS concepts were mapped to MHs or SCRs by pyMeSHSim directly. Meanwhile, non-MeSH-synonymous UMLS concepts were processed as free texts into MeSH-synonymous UMLS concepts, and then mapped to MeSH terms. All gene symbols were mapped to Entrez IDs. SCRs were mapped to its broader MHs. MHs with at least 10 genes in at least two groups were remained for further analysis.

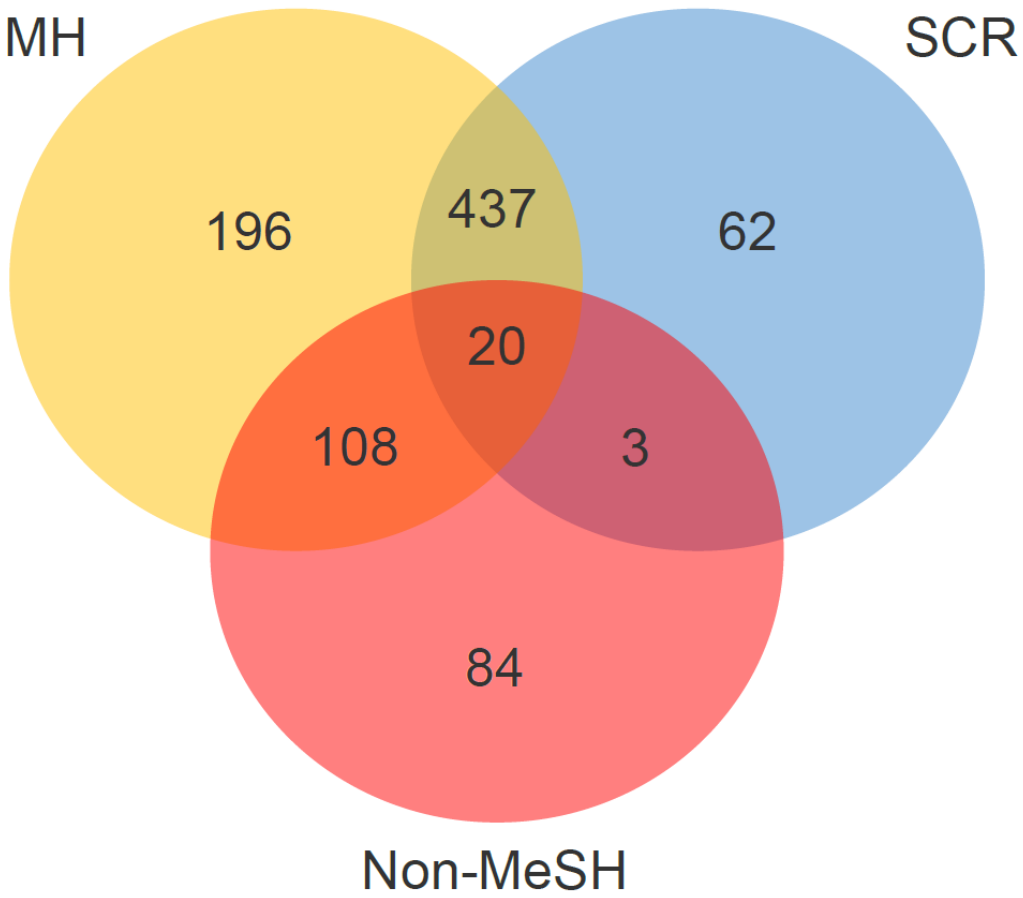


**Figure 3**. **Venn diagram of MH-gene pairs in MH, SCR and Non-MeSH groups**.