How to use MeSH-related Packages

Koki Tsuyuzaki[1,4], Gota Morota[2], Takeru Nakazato[3] and Itoshi Nikaido[4].

January 15, 2016

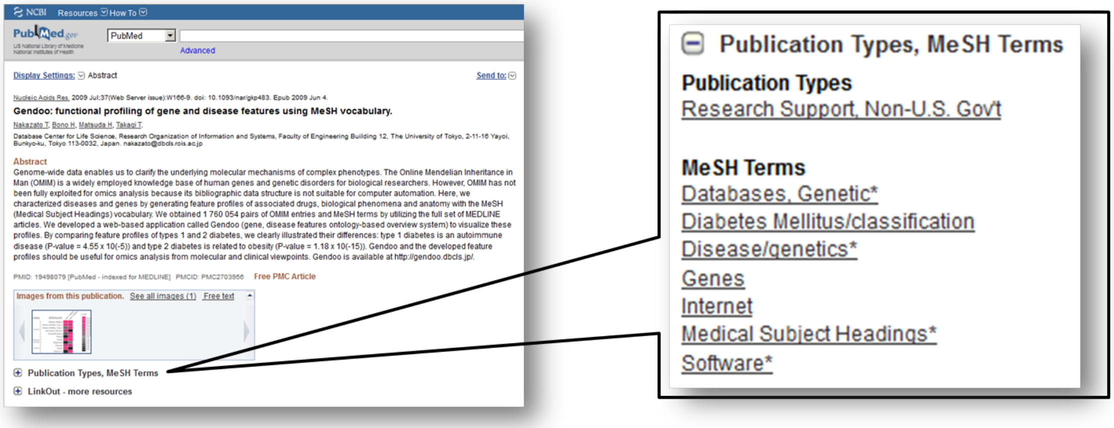
[1]Department of Medical and Life Science, Tokyo University of Science.
[2]Department of Animal Science, University of Nebraska-Lincoln
[3]Database Center for Life Science, Research Organization of Information and Systems.
[4]Bioinformatics Research Unit, RIKEN Advanced Center for Computing and Communication.

`k.t.the-answer@hotmail.co.jp, dritoshi@gmail.com`

Contents

1 Introduction

This document provides the way to use MeSH-related packages; *MeSH.db*, *MeSH.AOR.db*, *MeSH.PCR.db*, *MeSH.XXX.eg.db*-type packages, *MeSHDbi*, and *meshr* packages. MeSH (Medical Subject Headings) is the NLM (U. S. National Library of Medicine) controlled vocabulary used to manually index articles for MEDLINE/PubMed [1] and is a collection of a comprehensive life science vocabulary. MeSH contains more than 25,000 clinical and biological terms. The amount of MeSH term is about twice as large as that of GO (Gene Ontology)[2] and its categories are also wider. MeSH in 2014 proposed its 19 categories and *MeSH.db* provides 16 of them, which are actually assigned to some MeSH terms. Each category is expressed as single capital alphabet as abbreviation defined by NLM. Therefore MeSH is an expected to be much detailed and exhaustive gene annotation tool. Some software or databases using MeSH are now proposed [3, 4, 5, 6].


NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Advanced
Display Settings: Abstract
Send to:

Nucleic Acids Res. 2009 Jul;37(Web Server issue):W166-9. doi: 10.1093/nar/gkp483. Epub 2009 Jun 4.

Gendoo: functional profiling of gene and disease features using MeSH vocabulary.

Nakazato T, Bono H, Matsuda H, Takagi T.

Database Center for Life Science, Research Organization of Information and Systems, Faculty of Engineering Building 12, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan. nakazato@dbcls.rois.ac.jp

Abstract

Genome-wide data enables us to clarify the underlying molecular mechanisms of complex phenotypes. The Online Mendelian Inheritance in Man (OMIM) is a widely employed knowledge base of human genes and genetic disorders for biological researchers. However, OMIM has not been fully exploited for omics analysis because its bibliographic data structure is not suitable for computer automation. Here, we characterized diseases and genes by generating feature profiles of associated drugs, biological phenomena and anatomy with the MeSH (Medical Subject Headings) vocabulary. We obtained 1 760 054 pairs of OMIM entries and MeSH terms by utilizing the full set of MEDLINE articles. We developed a web-based application called Gendoo (gene, disease features ontology-based overview system) to visualize these profiles. By comparing feature profiles of types 1 and 2 diabetes, we clearly illustrated their differences: type 1 diabetes is an autoimmune disease (P-value = 4.55 x 10(-5)) and type 2 diabetes is related to obesity (P-value = 1.18 x 10(-15)). Gendoo and the developed feature profiles should be useful for omics analysis from molecular and clinical viewpoints. Gendoo is available at http://gendoo.dbcls.jp/.

PMID: 19498079 [PubMed - indexed for MEDLINE] PMCID: PMC2703956 Free PMC Article

Publication Types, MeSH Terms

Publication Types

Research Support, Non-U.S. Gov't

MeSH Terms

Databases, Genetic*
Diabetes Mellitus/classification
Disease/genetics*
Genes
Internet
Medical Subject Headings*
Software*

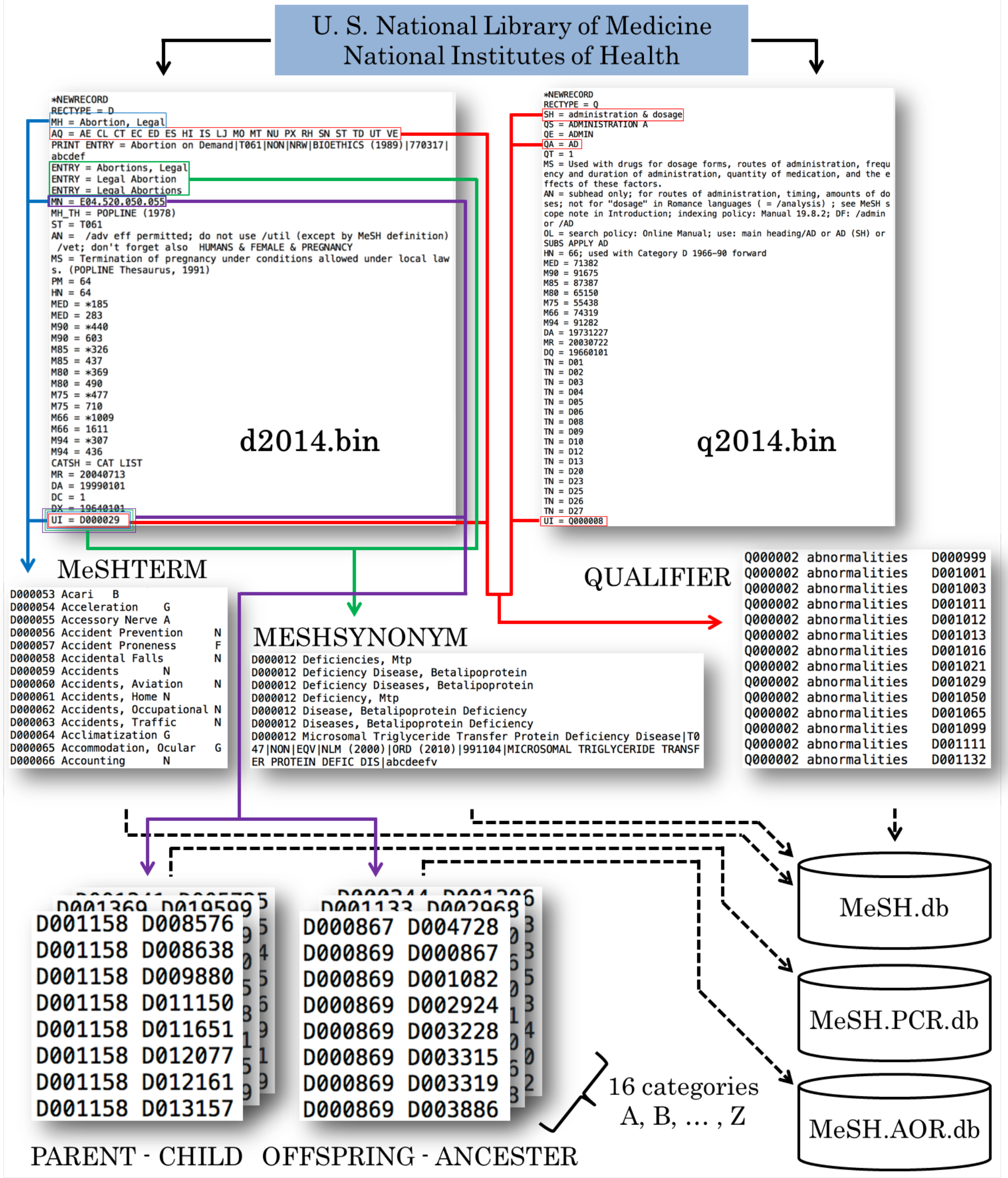
Figure 1: MeSH Term

This vignette introduces R/Bioconductor packages for handling MeSH in R. Original MeSH data is accessible by NLM FTP site (`http://www.nlm.nih.gov/mesh/filelist.html`). The data are downloadable as plain-text format (ASCII MeSH; d2015.bin / q2015.bin). These files were pre-processed by our data-processing pipeline (figure 2) and corresponding information is summarized as a table in SQLite3 file and packed into *MeSH.db*, *MeSH.AOR.db*, and *MeSH.PCR.db*.

Abbreviation	Category
A	Anatomy
B	Organisms
C	Diseases
D	Chemicals and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology and Food and Beverages
K	Humanities
L	Information Science
M	Persons
N	Health Care
V	Publication Type
Z	Geographical Locations

1.1 About MeSH

MeSH.db provides the corresponding table which contains MeSH ID, MeSH term, MeSH category, synonym, qualifier ID, and qualifier term. Qualifier term means more rough annotation (subheadings) than MeSH. MeSH has hierarchical structure like GO. Such structure is provided as *MeSH.AOR.db* (AOR: ancestor-offspring Relationships) and *MeSH.PCR.db* (PCR: parent-child Relationships as corresponding table.



```
*NEWRECORD
RECTYPE = D
MH = Abortion, Legal
AQ = AE CL CT EC ED ES HI IS LJ MO MT NU PX RH SN ST TD UT VE
PRINT ENTRY = Abortion on Demand|T061|NON|NRW|BIOETHICS (1989)|770317|
abcdef
ENTRY = Abortions, Legal
ENTRY = Legal Abortion
ENTRY = Legal Abortions
MN = E04.520.050.055
MH_TH = POPLINE (1978)
ST = T061
AN =  /adv eff permitted; do not use /util (except by MeSH definition)
 /vet; don't forget also  HUMANS & FEMALE & PREGNANCY
MS = Termination of pregnancy under conditions allowed under local law
s. (POPLINE Thesaurus, 1991)
PM = 64
HN = 64
MED = *185
MED = 283
M90 = *440
M90 = 603
M85 = *326
M85 = 437
M80 = *369
M80 = 490
M75 = *477
M75 = 710
M66 = *1009
M66 = 1611
M94 = *307
M94 = 436
CATSH = CAT LIST
MR = 20040713
DA = 19990101
DC = 1
DX = 19640101
UI = D000029
```

```
*NEWRECORD
RECTYPE = Q
SH = administration & dosage
QS = ADMINISTRATION A
QE = ADMIN
QA = AD
QT = 1
MS = Used with drugs for dosage forms, routes of administration, frequ
ency and duration of administration, quantity of medication, and the e
ffects of these factors.
AN = subhead only; for routes of administration, timing, amounts of do
ses; not for "dosage" in Romance languages ( = /analysis) ; see MeSH s
cope note in Introduction; indexing policy: Manual 19.8.2; DF: /admin
or /AD
OL = search policy: Online Manual; use: main heading/AD or AD (SH) or
SUBS APPLY AD
HN = 66; used with Category D 1966-90 forward
MED = 71382
M90 = 91675
M85 = 87387
M80 = 65150
M75 = 55438
M66 = 74319
M94 = 91282
DA = 19731227
MR = 20030722
DQ = 19660101
TN = D01
TN = D02
TN = D03
TN = D04
TN = D05
TN = D06
TN = D08
TN = D09
TN = D10
TN = D12
TN = D13
TN = D20
TN = D23
TN = D25
TN = D26
TN = D27
UI = Q000008
```

```
D000053 Acari   B
D000054 Acceleration    G
D000055 Accessory Nerve A
D000056 Accident Prevention     N
D000057 Accident Proneness      F
D000058 Accidental Falls        N
D000059 Accidents       N
D000060 Accidents, Aviation     N
D000061 Accidents, Home N
D000062 Accidents, Occupational N
D000063 Accidents, Traffic      N
D000064 Acclimatization G
D000065 Accommodation, Ocular   G
D000066 Accounting      N
```

```
D000012 Deficiencies, Mtp
D000012 Deficiency Disease, Betalipoprotein
D000012 Deficiency Diseases, Betalipoprotein
D000012 Deficiency, Mtp
D000012 Disease, Betalipoprotein Deficiency
D000012 Diseases, Betalipoprotein Deficiency
D000012 Microsomal Triglyceride Transfer Protein Deficiency Disease|T0
47|NON|EQV|NLM (2000)|ORD (2010)|991104|MICROSOMAL TRIGLYCERIDE TRANSF
ER PROTEIN DEFIC DIS|abcdeefv
```

```
Q000002 abnormalities   D000999
Q000002 abnormalities   D001001
Q000002 abnormalities   D001003
Q000002 abnormalities   D001011
Q000002 abnormalities   D001012
Q000002 abnormalities   D001013
Q000002 abnormalities   D001016
Q000002 abnormalities   D001021
Q000002 abnormalities   D001029
Q000002 abnormalities   D001050
Q000002 abnormalities   D001065
Q000002 abnormalities   D001099
Q000002 abnormalities   D001111
Q000002 abnormalities   D001132
```

```
D001158 D008576
D001158 D008638
D001158 D009880
D001158 D011150
D001158 D011651
D001158 D012077
D001158 D012161
D001158 D013157
```

```
D000867 D004728
D000869 D000867
D000869 D001082
D000869 D002924
D000869 D003228
D000869 D003315
D000869 D003319
D000869 D003886
```

Figure 2: Data pre-process for MeSH.db

1.2 The correspondence between MeSH ID and NCBI Entrez Gene ID

MeSH.XXX.eg.db (XXX is an abbreviation of species name such as Hsa: Homo sapiens) packages provide the correspondence between Entrez Gene IDs and NLM MeSH IDs. Such correspondence in wide variety of organisms are summarized as each MeSH.XXX.eg.db by three way of methods, Gendoo[4], gene2pubmed, and RBBH (reciprocal BLAST best Hit).

Gendoo is the web-application based on text-mining of PubMed. Co-occurrence relations in PubMed document are exhaustively retrieved and much relevant correspondence are filtered by some information science techniques.

gene2pubmed is the correspondence between Entrez Gene IDs and NLM PubMed IDs. These relationship is manually assigned by NCBI curator teams. We also summarized the relationship between MeSH Terms and PubMed IDs from licensed-PubMed, then merged as Gene IDs - MeSH IDs correspondence.

For some minor species including non-model organisms, which have no sufficient databases for annotation, we defined 15 well-annotated organisms and 100 minor-organisms, then conducted RBBH between all possible combinations using BLASTP search.

Method	Way of corresponding Entrez Gene IDs and MeSH IDs
Gendoo	Text-mining
gene2pubmed	Manual curation by NCBI teams
RBBH	sequence homology with BLASTP search (E-value $< 10^{-50}$)

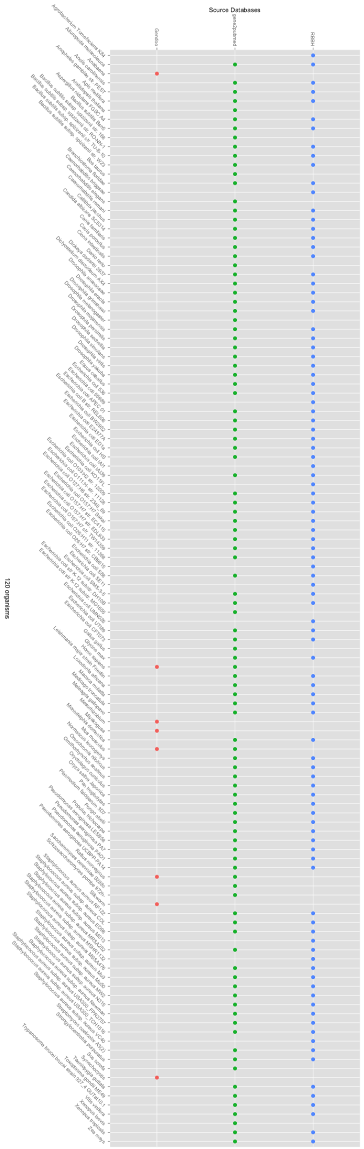


Figure 3: 120 organisms for MeSH.XXX.eg.db and those source databases

1.3 Database interface package for MeSH-related packages

We also implemented a database interface (DBI) package named *MeSHDbi*. This package is important because of two reasons. First reason is a unification of DBI functions for MeSH-related packages. *MeSH.db*, *MeSH.AOR.db*, *MeSH.PCR.db*, and *MeSH.XXX.eg.db* packages inherit the MeSHDb-class defined by *MeSHDbi* and behavior of these packages is uniformly designed. Second reason is supporting construction of user's original MeSH.XXX.eg.db package. Due to the rapid development of DNA sequence technology, wide variety of genome sequences are more and more determined and the correspondence of Gene IDs and MeSH IDs may be designed by many databases [3, 4, 5, 6]. Therefore, we prepared the function to create MeSH.XXX.eg.db package for a situation in which users can retrieved the relationship between Gene IDs and MeSH IDs by some means.

References

[1] S. J. Nelson and et al. The MeSH translation maintenance system: structure, interface design, and implementation. *Stud. Health Technol. Inform.*, 107: 67-69, 2004.

[2] M. Ashburner and et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1): 25-29, 2000.

[3] T. Nakazato and et al. BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes. *In Silico Biol.*, 8(1): 53-61, 2007.

[4] T. Nakazato and et al. Nucleic Acids Res. *Gendoo: functional profiling of gene and disease features using MeSH vocabulary.*, 37: W166-W169, 2009.

[5] D. J. Saurin and et al. GeneMeSH: a web-based microarray analysis tool for relating differentially expressed genes to MeSH terms. *BMC Bioinformatics*, 11: 166, 2010.

[6] M. A. Sartor and et al. Metab2MeSH: annotating compounds with medical subject headings. *Bioinformatics*, 28(10): 1408-1410, 2012.