Integracija podataka o asociranosti gen-bolest iz različitih izvora u cilju kreiranja anotacijskog fajla za Ontologiju bolesti

Sadržaj

Cilj

Postoji veliki broj baza podataka, koje sadrže informacije o povezanosti gena i bolesti. Međutim, ove baze čuvaju informacije u formatima, koji su često međusobno nekompatibilni, nisu uvek jednostavni za pretragu ili nisu zgodni za računarsku obradu. S druge strane, postoji veliki broj informacija koje se nalaze u naučnim radovima, a nisu prisutne u trenutno postojećim bazama. Ekstrakcija ovih “zarobljenih” informacija uveliko bi doprinela boljem razumevanju genske osnove mnogih patoloških stanja. Cilj master rada je kreiranje anotacijskog fajla koji bi sadržavao informacije o asociranosti gena i bolesti, s tim da bi svaka bolest bila predstavljena DO identifikacijom. Informacije bi bile prikupljane iz baza podataka kao što su DisGenNet, ClinVar, HumsVar, CDT (The Comparative Toxicogenomics Database), kao i metodom “text

mining” iz apstrakata dostupnih u PUBMED-u. U nastojanju da se cilj ostvari biće neophodno kreiranje programa koji će biti u stanju da prikuplja i integriše podatke iz različitih izvora.

Izvori informacija

Načini dobijanja informacija

Informacije o povezanosti bolesti sa odogvarajućim genom možemo da dobijemo na dva načina: pretragom specijalizovanih baza podataka i metodom "text mining".

Kriterijumi za odabir baze podataka

Ključni kriterijum za odabir baze podataka je njena ažuriranost. Baza koja nije ažurirana više od godinu dana ne uzima se u obzir, prilikom odabira. Drugi važan kriterijum je licenca pod kojom se podaci u bazi podataka nalaze. Uzimaju se u obzir samo baze podataka sa otvorenom licencom ili baze koje dozvoljavaju korišćenje podataka u nekomercijalne svrhe.

Načini pristupa podacima

Većina baza pohranjuje relevatne informacije o asociranosti gen-bolest u datoteke, koje su formatirane kao .tsv, .csv ili .txt. Ove datoteke se na jednostavan način mogu preuzimati. Neke baze podataka nude i odgovarajući API (engl. Application Programming Interface).

Odabir relevatnih podataka

Datoteke koje potiču iz različitih izvora mogu dosta varirati u količini informacija koje nude korisniku. Ove informacije možemo podeliti na dve grupe: a.) esencijalne ili ključne i b.) pomoćne. Esencijalne informacije su one koje se direktno odnose na gen i odgovarajuću bolest prouzrokovanu promenama na datom genu. Pomoćne informacije mogu da definišu tačan položaj gena u genomu, izvor odakle je asocijacija gen-bolest preuzeta, tip tkiva koji bolest pogađa, OMIM identifikacija (OMIM ID) i tome slično. Neke od ovih informacija kao što je izvor su bitne za predmet istraživanja, dok je recimo tačan položaj gena u genomu podatak koji može da se zanemari. S obzirom na prethodno rečeno, iz datoteka bi trebalo preuzeti identifikaciju gena, naziv bolesti i izvor odakle je informacija preuzeta i OMIM ID ako je prisutna. Ova identifikacija je bitna, jer na osnovu nje možemo automatski određenoj bolesti dodeliti identifikaciju Ontologije bolesti (DOID).

Odabrane baze podataka

S obzirom na ranije pomenute kriterijume odabrane su sledeće baze podataka: DisGeNet, CTD, ClinVar, Cosmic, HumsVar, HPO, Decipher, Orphanet, OMIM, GWAS catalog. Važno je napomenuti da ova lista nije konačna i može da se proširi.

Text mining

U cilju pronalaženja asocijacija gen-bolest, koje se ne nalaze u dostupnim bazama podataka, koristiće se metoda "text mining". Za potrebe istraživanja izabrana je specijalizovana alatka Diseases, koja može da se koristi kao *online* ili desktop aplikacija. Diseases je program otvorenog koda i kao takav je pogodan za integraciju u širi *framework*.

Anotacijski fajl

Struktura anotacijskog fajla

Anotacijski fajl bi trebalo da ima sedam polja odvojenih tabulatorima (pogledati tabelu dole).

Gene Symbol	Entrez ID	Uniprot ID	Ensembl ID	DOID	Sources	Disease name
PAX3	5077	P23760	ENSG00000135903	4051	cosmic	Alveolar rhabdomyosarcoma
PAX3	5077	P23760	ENSG00000135903	4051	clinvar	Alveolar rhabdomyosarcoma

Identifikacije gena

Prva četiri polja u gornjoj tabeli namenjena su identifikacijama gena. Ove identifikacije mogu da potiču od datoteke koja je preuzeta sa neke od baze podataka ili može biti naknadno dodata. Zapravo, u većini slučajeva datoteke koje preuzimamo i koristimo za kompilaciju, neće imati sve četiri identifikacije. U najvećem broju slučajeva biće prisutan Gene Symbol i Entrez ID. Ovo i nije toliko veliki problem, jer potreban je samo jedna vrsta identifikacije da bi se dobile ostale. Proces kojim na osnovu jedne vrste identifikacije dobijamo drugu naziva se mapiranjem. Postoji nekoliko resursa na internetu, kao što su Uniprot i HUGO, koji redovno održavaju datoteke sa informacijama o povezanosti različitih identifikacija gena. Ove datoteke mogu slobodno da se preuzimaju.

Identifikacija Ontologije bolesti

Peto polje u tabeli namenjeno je identifikacije Ontologije bolesti (DOID). Ontologija bolesti svakoj grupi oboljenja ili pojedinačnom oboljenju dodeljuje jedinstvenu identifikaciju. DOID ima sledeći oblik DO:[0-9]+ . Informacija o DOID-u koji odgovara određenoj bolesti može da se dobije iz ulaznog fajla (fajl preuzet sa neke od baza podataka), da se dobije indirektno na osnovu OMIM identifikacije ili da se ručno dodeli ako to nije moguće na prethodna dva načina. Dodeljivanje DOID-a na osnovu OMIM identifikacije zahteva dodatno objašnjenje. Ontologija bolesti može da se preuzme u formi OBO fajla. U fajlu je svakom terminu dodeljeno odgovarajuće ime i DOID. U nekim slučajevima je DOID direktno mapiran na odgovarajući OMIM ID, kao u primeru datom dole.

```
[Term]
id: DOID:0050167
name: autoimmune polyendocrine syndrome type 1
def: "An autoimmune polyendocrine syndrome that is inherited in an autosomal
recessive fashion, which is characterized by abnormal functioning of the immune
system that causes auto-reactivity against endocrine organs."
[url:https\://rarediseases.info.nih.gov/diseases/8466/autoimmune-polyglandular-
syndrome-type-1] {comment="sn:IEDB"}
synonym: "autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy" EXACT
[]
synonym: "Autoimmune Polyglandular Syndrome I" EXACT []
synonym: "Whitaker syndrome" EXACT []
xref: GARD:8466
xref: OMIM:240300
is_a: DOID:14040 ! autoimmune polyendocrine syndrome
```

Izvori

Šesto polje u anotacijskom fajlu namenjeno je za izvore informacija o asocijacije gena i bolesti. U slučaju da jedna asocijacija ima više izvora, onda se za svaki izvor asocijacija gen-bolest stavlja u poseban red, kao što je slučaj u gore prikazanoj tabeli.

Naziv bolesti

Poslednja kolona sadrži naziv bolesti ili grupe oboljenja.

BAZE PODATAKA

1. baza: DisGeNet

naziv fajla: curated_gene_disease_associations

link: https://www.disgenet.org/downloads (curated fajl)

sta imamo: gene symbol, disease name

sta nedostaje: Entrez ID, Uniprot ID, Ensembl ID, DOID

2. baza: Cosmic
link: https://cancer.sanger.ac.uk/cosmic/download
naziv fajla: cancer_gene_census.csv
napomene:
- postoje tri linka za download - whole, filtered i scripted. Biramo whole.
- Za naziv bolesti uzimamo informacije iz sledece dve kolone: Tumour Types(Somatic) i Tumour Types(Germline). Ukoliko je navedeno vise bolesti, svaku treba navesti u posebnom redu u anotacijskom fajlu.
- neophodna registracija mejlom sa akademske mreze (alas)

3. baza: HumsVar
link: https://www.uniprot.org/docs/humsavar
napomena:
- Ovde se uzimaju u obzir samo redovi kood kojih je Type of variant=Disease (jedino je za njih dato Disease name)

sta imamo: gene symbol, disease name
sta nedostaje: Entrez ID, Uniprot ID, Ensembl ID, DOID

4. Orphanet
link: http://www.orphadata.org/cgi-bin/index.php
naziv fajla: http://www.orphadata.org/data/xml/en_product6.xml

sta imamo: gene symbol, disease name, Ensembl ID
sta nedostaje: Entrez ID, Uniprot ID, DOID

5. ClinVar
link: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/
naziv fajla: gene_condition_source_id

sta imamo: gene symbol, disease name
sta nedostaje: Entrez ID, Uniprot ID, Ensembl ID, DOID

6. HPO
link: https://hpo.jax.org/app/download/annotation

naziv fajla: genes_to phenotypes.txt
(http://compbio.charite.de/jenkins/job/hpo.annotations/lastSuccessfulBuild/artifact/util/annotation/genes_to_phenotype.txt)

sta imamo: Entrez ID,disease-ID for link
sta nedostaje: gene symbol, Uniprot ID, Ensembl ID, DOID, disease name

7. Diseases (pretraga naucne literature)
link: https://diseases.jensenlab.org/Downloads
naziv fajla: human_disease_textmining_filtered.tsv (Text mining channel filtered)

sta imamo: gene symbol, DOID, disease name
sta nedostaje: Entrez ID, Uniprot ID, Ensembl ID

napomena:

- Potrebno je proveriti da li je DOID identifikator mozda zastareo u dokumentu na sledecem linku: http://www.obofoundry.org/ontology/doid.html. Ako je is_obsolete true, znaci da jeste i da ga ne treba ukljuciti u anotacijski fajl

  [Term]
  id: DOID:0050001
  name: obsolete Actinomadura madurae infectious disease
  subset: gram-positive_bacterial_infectious_disease
  is_obsolete: true

~~ Ocekuje se dodavanje jos nekoliko baza

MAPIRANJE

To su baze iz kojih dobijamo informacije o gene-disease vezama. Dalje, preostale informacije iz tabele dobijamo:

1. Uniprot (za informacije o genima)
   link:
   ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/
   naziv fajla: HUMAN_9606_idmapping.dat

primer:

P23760 UniProtKB-ID PAX3_HUMAN
P23760 Gene_Name PAX3

P23760 Gene_Synonym HUP2
…
P23760 GeneID 5077
...

Imamo gene_name (PAX3), Uniprot_ID (P23760), EntrezID (5077), za informacije o genima nedostaje EnsemblID, to pronalazimo u HUGO datoteci

2. HUGO (za informacije o genima)

- potrebno je u formularu https://www.genenames.org/download/custom/ izabrati nekoliko polja, npr: Approved symbol, NCBI Gene ID, Approved name, Chromosome, Ensembl gene ID, Pubmed IDs, CCDS IDs, Status, Accession numbers, RefSeq IDs - gene name moze nekad da se nadje u approved symbol pa na osnovu toga da se dobije Ensembl gene ID

3. .OBO fajl (za informacije o bolestima)

link: http://www.obofoundry.org/ontology/doid.html

[Term]
id: DOID:0050001
name: obsolete Actinomadura madurae infectious disease
subset: gram-positive_bacterial_infectious_disease
is_obsolete: true

Ako je poznato ime bolesti, mozemo dobiti DOID