PGxCorpus - Annotation guidelines

December 20, 2017

Contents

1 Purpose

This document presents how to manually and homogeneously annotate the PGx-Corpus. First it presents the annotation schema, *i.e.*, the types of entities and relationships to annotate. Secondly, it presents the general rules that may guide the manual annotation.
This document is derived from guidelines previously established for copora, such as the DDI Corpus of the Task 9 of Semeval 2013 [2] and the MERLOT corpus [1].

2 Annotation schema

2.1 Entities

Figure 1 presents the 10 types of entities to annotate. Entites of these types are automatically pre-annotated, then proposed to annotators. Annotators will correct and complete pre-annotations. Definitions of entity types and examples of annotations are provided in this subsection.

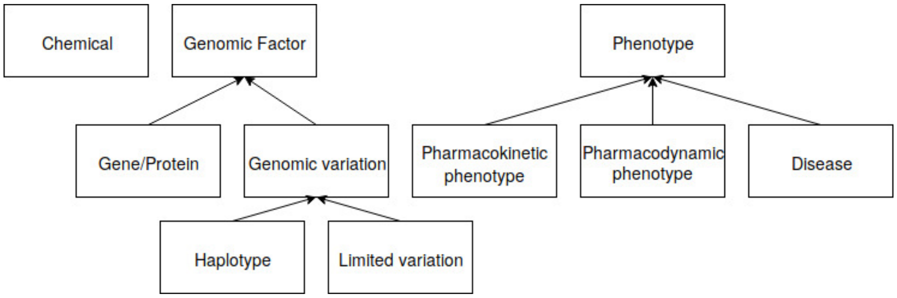


Figure 1: Types of entities to annotate in the PGxCorpus

2.1.1 Chemical

Chemicals are **any distinct compounds or substances** in particular, but not only, therapeutic agents (*i.e.*, drugs). Mentions of chemicals can be any molecule names (*e.g.*, 4-acetamidophenol), names of drug active ingredients (*e.g.*, acetaminophen), drug brand name (*e.g.*, Actimol) or classes of drugs (*e.g.*, anticoagulants).
For this corpus, we propose to annotate only chemicals **used in a clinical setting**, which means that any mention of chemicals naturally present in the human body should not be annotated if they has not been used especially for a therapeutic purpose.

Examples

- (Annotate molecule names): 2,2-Bis(4-hydroxyphenyl)propane, Adenosine triphosphate; 3-Methylmorphin

- (Annotate drug ingredients): omeprazole; codeine; sildenafil
- (Annotate drug brands): Actimol; Mopral; Viagra
- (Annotate drug classes): anticoagulant; anti-HIV treatment; chemotherapy

- (do NOT annotate endogenous chemicals): endogenous steroids, endogenous insulin

Unusual cases

In the case of **mixture of chemical**, all constituents of the mixture must be included in the annotation.

- (Annotate as one chemical): ...***Acetaminophen, Caffeine & 8mg Codeine Phosphate*** Caplets may be associated with...
- (do NOT annotate as several chemicals): ...***Acetaminophen***, ***Caffeine*** & 8mg ***Codeine Phosphate*** Caplets may be associated with...

Food should not be annotated as chemical.

- (do NOT annotate as a chemical): brocoli; grapefruit

In the case of **ambiguous** entities that may refers both to a chemical and to another entity, as a gene/protein name (*e.g.*, insulin) or food (*e.g.*, alcohol), then the context of the sentence will be used to clarify if the entity is used in a clinical setting or not. If the ambiguity can not be clarified with the sentence, then the sentence must be discarded.

- (Annotate as a chemical): ...***insulin*** treatment is impacted by...
- (do NOT annotate as a chemical): ...***insulin*** gene expression is inhibited by...

Links to search in DrugBank and PubChem are provided and may help when doubts about the classification of an entity as chemical appears.

2.1.2 Genomic factor

Genomic factors are any **elements of the genome** such as chromosomes, genes (and by extension gene products, *i.e.*, proteins), exons, simple or complex genomic variations.

Any genomic factor that does not fit with the sub-types described below should be annotated broadly as a genomic factor.

- (Annotate): CYP2D6 genotype

Gene (and protein)

Mentions of genes and their products, proteins, must be annotated in text. They may be mentioned in the text with their official gene symbol (or protein name), a synonym or their full name. Family of gene/protein will also be annotated with this type.

- (Annotate): Thiopurine S-methyltransferase; TPMT; MG115; CYP2D6; cytochrome P450 2D6; cytochrome P450; kinase

Genomic variation

Genomic variations are elements of the genome that change between individuals or populations. These variations spread from large structural variation such as chromosomal duplications, Copy Number Variation (CNV), or limited sequence variations such as Single Nucleotide Polymorphisms (SNP). Any genomic variation that does not belong to one of the sub-types described below (Haplotype and Limited variation) should be annotated broadly as a Genomic variation.

- (Annotate): copy number variants of CYP2D6; mobile element insertion in CYP2C19; tandem duplication upstream to TPMT

Haplotype

An haplotype is a set of genetic variations, usually SNPs, inherited together that consequently tend to appear together. Haplotype names are frequently composed of a gene name and a suffix, such as in *CYP2C9* ***2**. If an haplotype is involved in a relation and contains a gene entity, the haplotype should be the entity pointed out in the relation.

- (Annotate): TPMT*3A; CYP2D6*5; TPMT*1; CYP2D6UM

Limited variation

These are inter-individual variations in genetic (or protein) sequences, such as SNP, In/del, mutations, etc. Annotators must annotate mentions of variants that may be very general, such as “polymorphisms allele/genotype of the gene VKORC1”, but also very precise, such as an *rs ids* issued from dbSNP. When provided, the localization of the mutation should always be included in the annotation.

- (Annotate): TPMT polymorphisms; rs7295; G6PD mutation; TPMT A719G; VKORC1 variants; NC_012920.1:m.1555A>G; wild-type TPMT allele; VKORC1 genotype; C/G genotype of rs7295; allele G of the CYP2D9 gene; Asp(9)Asn mutation in the lipoprotein lipase gene

2.1.3 Phenotype

Phenotype entities are **any observable characteristics** of an individual resulting from the interaction of its genotype with the environment. Please not that we consider as phenotypes, characteristics that may be observed at the clinical level (such as a symptom) as well as those observed at the molecular level (such as the activity or the concentration of an enzyme). We propose three sub-types of phenotypes: pharmacokinetic, pharmacodynamic phenotypes and diseases. Any phenotype that does not fall in one of these sub-types is to annotate with the broader type Phenotype.

- (Annotate): TPMT activity; Drop in neutrophils; hepatotoxicity; TPMT deficiency; myostatin transcription

Pharmacokinetic phenotype

Pharmacokinetics is the study of drug absorption, distribution, metabolism and excretion. Any observation of such action of the body on a drug is a pharmacokinetic phenotype.

- (Annotate): Hydroxylation of nortriptyline; Concentration of paroxetin; Warfarin disposition; S-methylation of 6-thiopurine;
- (Example): ...A ***warfarin requirement*** of < 2.5 mg/day and an elevated ***warfarin S/R concentration***...
  Here are two annotations of Pharmacokinetic phenotypes. In addition, each occurrence of ***warfarin*** must be annotated as a Chemical.

Pharmacodynamic phenotype

Pharmacodynamics is the study of the biochemical and physiological effects of drugs and the mechanisms of their actions. Any observation of a drug effect or action is a pharmacodynamic phenotype.

- (Annotate): Anticoagulant effect of warfarin; Response to mercaptopurine; Activity of thiopurine; Paroxetine-induced conversion of cytochrome P450 2D6

Disease

"A disease is a condition of the body that impairs normal functioning and is typically manifested by distinguishing signs and symptoms."
[`https://www.merriam-webster.com/dictionary/disease`]

- (Annotate): Acute lymphoblastic leukemia; Inflammatory bowel disease; Leukopenia

Diseases should be distinguished from symptoms: "a symptom is a physical or mental feature which is regarded as indicating a condition of disease" [`https://en.oxforddictionaries.com/definition/symptom`]. Symptoms must be annotated with the broader type Phenotype reither than

Disease. In the case of diseases that are symptoms of more complex diseases, they must be annotated as diseases.

- (do NOT annotate as Disease, but as Phenotype): fever; arrhythmia; rash

In case of doubt, please use the links to search on Malacards to distinguish.

Unusual case

In some articles, **population-level observations** are reported rather than individual phenotypes. If this population-level observation is associated with a phenotype name, then only the phenotype must be annotated, but if no proper phenotype accompanies the population-level observation, then this one must be annotated.

- (Example 1): ...Patients present an increased risk of **bleeding**...
  In this case, only bleeding must be annotated, not "risk of".
- (Example 2): ...Prevalence of **heart attack** in Asian populations...
  Here, only heart attack must be annotated.
- (Example 3): CXCR4 polymorphism predicts **progression-free survival** in metastatic colorectal cancer patients
  In this particular case, "progression-free survival" must be annotated because this population-level observation is not associated to a proper phenotype in the sentence.

This last case frequently occurs for mentions of overall survival (OR), overall response rate (ORR) and progression-free survival (PFS).

2.2 Relationships

Figure 2 presents the 9 types of relationships to annotate between pairs of entities. We consider, then annotate in this study only binary relationships.

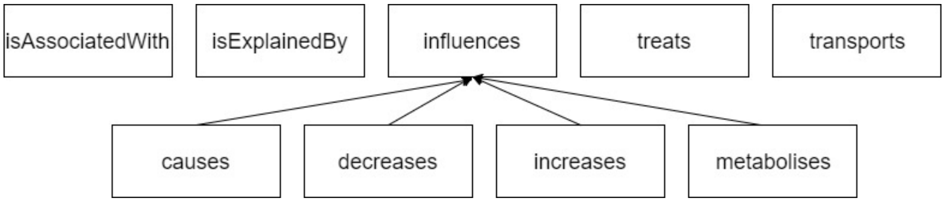
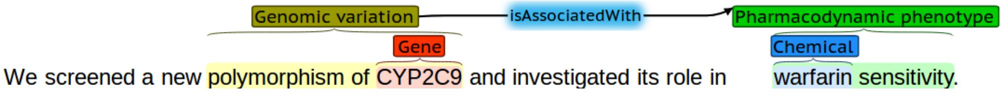


Figure 2: Types and sub-types of relationships to annotate in the PGxCorpus.

2.2.1 isAssociatedWith

IsAssociatedWith is the less precise and default type of relationships to bind two entities. It has no direction and is to be used when a relationships is mentioned, but no detail about the type of relationship is provided by the sentence, or when no other type is adapted.

Examples



We screened a new polymorphism of CYP2C9 and investigated its role in warfarin sensitivity.

“We studied the association between ***warfarin response*** and ***CYP2C9 poymorphisms*** in asian patients.”

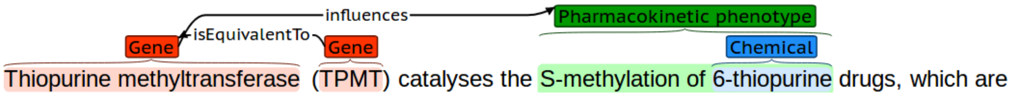
2.2.2 isExplainedBy

The observation of the first entity is explained by the second entity. In this case, the relationships is oriented.

“Interactions between ***warfarin response*** variation has been proven to be related to ***CYP2C9 polymorphism***.”

2.2.3 influences

The first entity affects or changes how behaves the second. The annotator should make sure that one of the sub-type of this relationship (*i.e.*, causes, metabolizes, increases, decreases) does not apply before annotating.



Thiopurine methyltransferase (TPMT) catalyses the S-methylation of 6-thiopurine drugs, which are

Examples

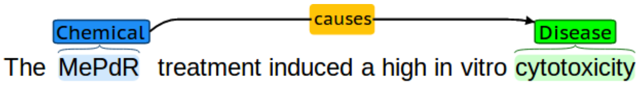
"The ***warfarin response*** is impacted by the level of ***expression of CYP2C9***."

"***Heroin analgesia*** is mediated by ***mu-opioid receptors***"

causes

The first entity is the reason why the second entity happen (usually a phenotype).

Example

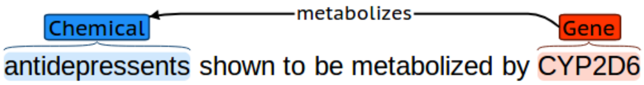


The MePdR treatment induced a high in vitro cytotoxicity

metabolizes

The first entity processes the metabolism of the second.

Example



antidepressents shown to be metabolized by CYP2D6

increases

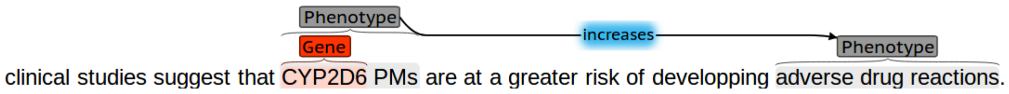
To (make something) become larger in amount or size.
[`http://dictionary.cambridge.org/dictionary/english/increase`]
If a first entity makes a second entity become larger in amount or size, the increases relationship must be used.

Example

decreases

To become less, or to make something become less.

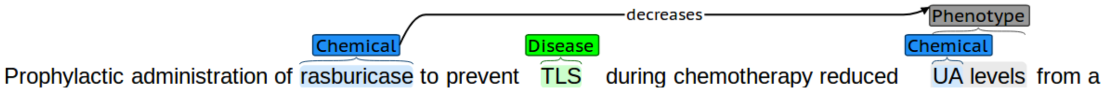
[`http://dictionary.cambridge.org/dictionary/english/decrease`]



clinical studies suggest that CYP2D6 PMs are at a greater risk of developping adverse drug reactions.

If a first entity makes a second entity become less, the decreases relationship must be used.

Examples



Prophylactic administration of rasburicase to prevent TLS during chemotherapy reduced UA levels from a

"In patients with ***BRCA1 mutations***, treatment with ***tamoxifen*** appears to reduce the risk of ***contralateral breast cancer*** development."

2.2.4 Transports

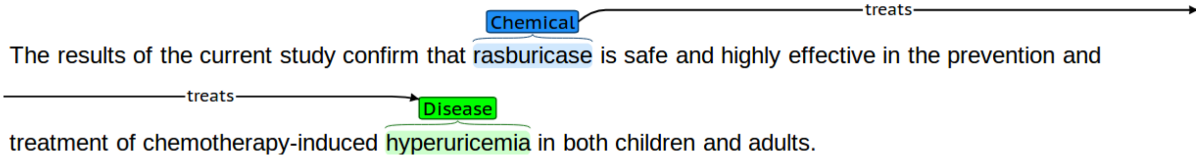
The first entity (usually a protein) transports the second (usually a chemical), for instance from one compartment to another.

Example
"Genetic variations in ***ABCB1*** affect the ***transport of codein***."

2.2.5 Treats

A drug treats a disease or a symptom.

Example



The results of the current study confirm that rasburicase is safe and highly effective in the prevention and treatment of chemotherapy-induced hyperuricemia in both children and adults.

"***Montelukast*** is a leukotriene receptor antagonist (LTRA) used for the maintenance treatment of ***asthma***"

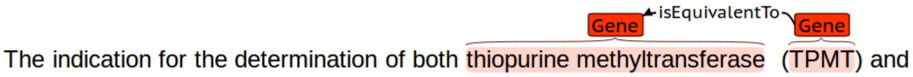
2.2.6 isEquivalentTo

An entity (often an abbreviation) has the same meaning that another entity.

Table 1: Values associates with each relationship attribute

attribute value	code
affirmed	- none -
negated	`neg`
hypothetically affirmed	`hyp`
hypothetically negated	`hyp neg`

Example



The indication for the determination of both thiopurine methyltransferase (TPMT) and

thiopurine methyltransferase and ***TPMT*** are two genes linked by the relationship isEquivalentTo.
If one needs to annotate another relationship involving equivalent entities, we ask annotators to choose only one of them and not to create one relationship for each of the equivalent entities. For consistency purposes we propose to annotators to choose the most relevant of the two entities by considering the sentence. Following the previous example, thiopurine methyltransferase (vs. TPMT) will be the entity to consider when annotating more relationships.

2.3 Relationship attributes

We propose to describe annotated relationships with an attribute:

- *Default* – A relationship is *affirmed* if the text clearly says that the two considered entities are related.

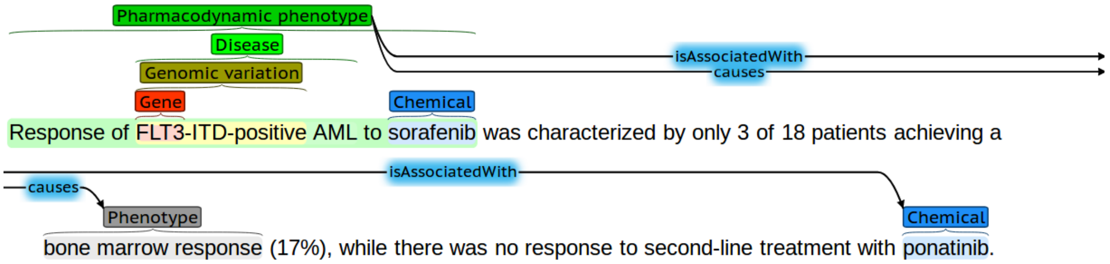
- `neg` – A relationship is *negated* if the text clearly says that the two entities are not related. Note that this is different from the absence of relationship, which may occur if nothing is mentioned about a potential relationships.

- `hyp` – A relationship is *hypothetically affirmed* if the text clearly mention that the mentioned relationship is uncertain, fuzzy, under study, or needs more evidences, but not negated.

- `hyp neg` – A relationship is *hypothetically negated* if the text clearly says that two entities are potentially not related.

To associate a relationship with its attribute, the annotator must report one of the three possible codes (`neg`, `hyp` or `hyp neg`) as a “Note” in the Brat web interface. Table 1 summarizes the codes associated with attribute values.

In this example, treatment with ponatinib triggered no response, which makes the isAssociatedWith relation negative. This relation is hypothetical



Response of FLT3-ITD-positive AML to sorafenib was characterized by only 3 of 18 patients achieving a bone marrow response (17%), while there was no response to second-line treatment with ponatinib.

Figure 3: Annotations where the relationship of type 'causes' is to associate with the attribute `hyp` (*hypothetically affirmed*) and the relationships of type 'isAssociatedWith' is to associate with the attribute `hyp neg` (*hypothetically negated*).

too as it is stated in the framework of a study on a group of patient.

Examples

"**TAU haplotype** does not influence the **treatment of Alzheimer's disease.**"
The TAU haplotype has a *negated* relationship with the drugs used to treat Alzheimer's disease.

"**Polymorphisms in the gene TMPT** may be involved in the **overreaction to thiopurines**"
TMPT variants are *hypothetically* related with thiopurine response.

"In this study , we evaluated whether **norcantharidin** exhibits **anticancer effects**..."
Norcantharidin is *hypothetically* related with anticancer effects.

3 Annotation rules

In order to create a consistent corpus, it is critical that annotators annotate sentences homogeneously. This section details how rules on how annotator must annotate entities and relationships they will find in texts.

1. **Consider context:** In any case, the context given by the sentence should be used in order to disambiguate any doubt about entities or relationships.

2. **Explicitness is needed:** Annotation of relationships should be done only if the relationship is explicitly mentioned in the text, particularly, no

inferred or guessed knowledge should lead to the annotation of a relationship.

3. **Stay within the framework:** Annotation should only be made if the entities and relationships involved are within the scope of the corpus, *i.e.*, pharmacogenomic relationships. Any sentence that is not related to pharmacogenomic should be discarded. For instance, any relationship that could occur between two Gene entities is (by definition) out of the scope of this corpus.

   - (example 1): The combination of **ACE inhibitors** with potassium-sparing **diuretics** such as **amiloride** can increase **potassium retention** so strongly that life-threatening **hyperkalemia** ensues
     This sentence is not about pharmacogenomic, but drug-drug interactions and is out of the scope of this corpus. It should be discarded regardless of the annotation that could be made in it (See subsection 4.4 on how to discard a sentence).
   - (example 2): CONCLUSIONS : We demonstrated that genetic variations in CCHCR1 are strongly associated with nevirapine -induced rash.
     This sentence is in the scope of this corpus. It can be annotated and should not be discarded.
     We can find 4 entities in it:
     (a) ***CCHCR1*** is a Gene.
     (b) ***genetic variations in CCHCR1*** is a Genomic variation.
     (c) ***nevirapine*** is a Chemical.
     (d) ***nevirapine-induced rash*** is a Phenotype.

     We can find one relation between the Genomic variation (genetic variations in CCHCR1) and the Phenotype (nevirapine-induced rash), of the 'isAssociatedWith' type.

   Sentences about genetical therapy, where the transfection of a gene may treat a disease should not be annotated, because this is not pharmacogenomics. Sentences about gene variation is the cause for a (Mendelian) genetic disease should be discarded, because this is not pharmacogenomics. Sentences about the fact that a drug may treat a disease should also be discarded, because this is not pharmacogenomics (even if it is pharmacology). However, *treats* relationships must be annotated within sentences about pharmacogenomics.

4. **Be specific:** The best precision is preferred. The annotation should include the largest part of text that describe an entity, since it will annotate the most specific entity. In the sentence "Finally , a novel **missense mutation** in the **perforin** was identified"

- (example 1): ...Finally , a novel **missense mutation** in the **perforin** was identified...
    - (annotate): missense mutation, perforin
    - (do NOT annotate): mutation
- (example 2): ...***genetic variations in CCHCR1***...
  In this sentence, ***CCHCR1*** should be annotated as a gene, and ***genetic variations in CCHCR1*** as a Genomic variation.
- (example 3): ...strongly associated with ***nevirapine-induced rash***.
  In this sentence, ***nevirapine*** should be annotated as a Chemical and ***nevirapine-induced rash*** should be annotated as a phenotype. "Nevirapine-induced" should be included in the span of the phenotype because it qualifies the rash.

5. **Stay consistent:** A minimal specificity is needed in the annotations. If the terms used are too common, they should not be annotated. The annotator will make decision whether an entity is too common to be annotated or not, referring to the following examples.

    - (example 1): ...renal nNOS gene expression...
      In this sentence, do NOT annotate renal. ***nNOS*** is a Gene. ***nNOS ... expression*** is a Phenotype.
      The annotator should annotate adjective, such as a localization adjective, only when this one takes part in the definition of the phenotype or disease.
    - (example 2): ...the CLPTM1L locus on susceptibility to lung cancer and sensitivity to...
      In this sentence, annotate ***lung cancer*** as a disease, because the localization participate in the definition of the disease.
    - (example 3): ...**Acute Lymphoblastic Leukemia** patients, treated with...
      In this sentence, annotate ***Acute Lymphoblastic Leukemia*** as a disease, because ALL is identified as a proper disease (see `http://www.malacards.org/card/leukemia_acute_lymphoblastic`).
    - (example 4): ...VKORC1 variants increase the risk of acute **bleeding**.../ Here, only bleeding should be annotated.
    - (example 5): ... MGMT polymorphisms impact the effect of chemotherapy treatments ...
      MGMT polymorphisms should be annotated as Limited variation, chemotherapy as Chemical, effect of chemotherapy as Phenotype. The word treatments is not to annotate.

6. **Discontinuous annotations:** If a single entity is mentionned by words spread in several parts of the sentence, these different parts should be annotated, using several fragments for a single entity.

7. **Skip co-references:** Co-reference should not be annotated. That means that pronouns should never be annotated.

    - (example): ...These SNPs affects the effect of chemotherapy ...
      These SNPs should not be annotated since it refers to polymorphisms described earlier in the article.

8. **Consider abbreviations:** The abbreviations should always be annotated as an entity in its own right and refer to its extended form when possible with the isEquivalentTo relationship.

9. **Multitasking entities:** An entity can be involved in several relationships at once and these relationships should all be annotated according to the guidelines.

10. **Take no risk:** If a sentence is ambiguous or mistakenly chosen for the annotation task, the annotator must discard it. Any typographic error that can lead to a misunderstanding of the sentence should lead to the discard of the sentence. How to discard a sentence is explain in subsection 4.4.

    - (example 1): that functional polymorphisms in FAS and FAS ligand (FASL) are associated with susceptibility to lung cancer and esophageal cancer.
      This sentence is uncompleted (it does not start at the beginning of the sentence) then must be discarded.
    - (example 2): ...the NAT enzyme metabolises inactivation of isoniazid...
      The verb metabolise is not appropriate since one cannot metabolise an inactivation

    Similarly, when the annotator doubts too much about the annotation, the sentence should be discarded.

11. **Add prepositions:** In entities which are described with multiple terms, prepositions should be annotated, whereas the pre-annotation tool does not include them.

    - (example 1): ...mutation in the perforin...
      In this sentence, ***perforin*** should be annotated as a Gene and ***mutation in the perforin*** should be annotated as a Genomic variation. Whereas, the pre-annotation tool only annotates mutation... perforin.
    - (example 2): ...substitution of Met for Ile...
      In this sentence, ***substitution of Met for Ile*** should be annotated as a Genomic variation.

- (example 3): ...adverse drug reactions toward mercaptopurines...
  In this sentence, ***mercaptopurines*** should be annotated as a Chemical and ***adverse drug reactions toward mercaptopurines*** should be annotated as a Phenotype.

12. **Annotate relations within nominal groups:** Relations may occurs within a nominal group and in this case must be annotated.

    - (example 1): ...**CYP2D6 variants-codeine** association ...
    - (example 2): ...**cocaine-induced phosphorilation** of glutamate receptors ...

4 Annotation tool: Brat

The tool proposed for the manual annotation task is Brat [3], a web-based and open source annotation tool. An automatic pre-annotation is performed on the texts to annotate using PubTator [4], plus house-made software. The pre-annotation only underline entities to facilitate annotator task, but they need to be corrected.
Some quick search links will be provided on Brat in order to ease the annotator's work, using the databases cited above.

4.1 Using Brat

Each annotators is provided with an access (login, password) to our Brat server (`https://pgxbrat.loria.fr/`) in order to annotate sentences. Brat is very intuitive. The main actions it enables are here described:

- Connexion: To get connected, the annotator has to go to `https://pgxbrat.loria.fr/#/login`, where `login` must be replaced by the annotator login. Then, a pop-up will appear, asking for credentials. Unique credentials will be provided by the administrator to each annotator.

- Annotate an entity: To annotate an entity, the annotator select a span of text with his mouse. When the left button of the mouse will be released, a pop-up window will appear, asking the annotator to select the type of entity the peace of text mention.

- Annotate a discontinuous entity: To annotate a discontinuous entity, the annotator will first have to annotate a part of the entity, then, double-click on the entity created and click on the "Add Frag." button on the bottom of the pop-up window.

  Similarly, the "Move" and "Move Frag." buttons enable to shift the part of text associated with an entity, or with the fragment of an entity.

- Annotate a relationship: To annotate a relationship, the annotator click on the first entity of the relationship and drag to the second entity. An arrow will then appear and a pop-up will appear, asking the annotator to select the type of relationship.

In case of question, consul the brat manual (`http://brat.nlplab.org/manual.html`) or email us at `pgxcoprus@inria.fr`.

4.2 Explore web resources to help annotating

Brat enable the annotator to search various resources for a span of text he wants to annotate. This feature may help to ensure a decision about annotating or not an element with a particular type:

4.2.1 Diseases

Malacards (`http://www.malacards.org/`) is a database about human **diseases**.
CTD (`http://ctdbase.org/`) is a database about toxicogenomics, which includes diseases related with environmental exposure.

4.2.2 Chemical

DrugBank (`https://www.drugbank.ca/`) is a database about drugs and can be used to ensure whether some words or group of words refer to a drug or not.
PubChem (`https://pubchem.ncbi.nlm.nih.gov/`) is a database about chemicals, in a broad sens. PubChem has a wider range than DrugBank, consequently please, ensure that the chemical you could find in this database may be used in a clinical purpose before to annotate it as a chemical (see section 4).

4.2.3 Genomic Factor

NCBI Gene (`https://www.ncbi.nlm.nih.gov/gene`) is a database about genes and proteins.
DbSNP (`https://www.ncbi.nlm.nih.gov/projects/SNP/`) provides information about single nucleotide polymorphisms.

4.3 Estimated time for annotating

We estimated that annotating one sentence with Brat should take about 4 minutes.

4.4 Discard a sentence

If a sentence is out of the scope of PGxCorpus or presents obvious problems (incompletness, typos, ambiguity, etc.) the annotator must discard it. To discard a sentence, an annotator needs to remove all annotations from the sentence, leaving it as a simple plain sentence.

5 Any question?

Questions, such as issues to connect to Brat or to understand these guidelines, are to email at `pgxcorpus@inria.fr`.

6 Examples of annotation

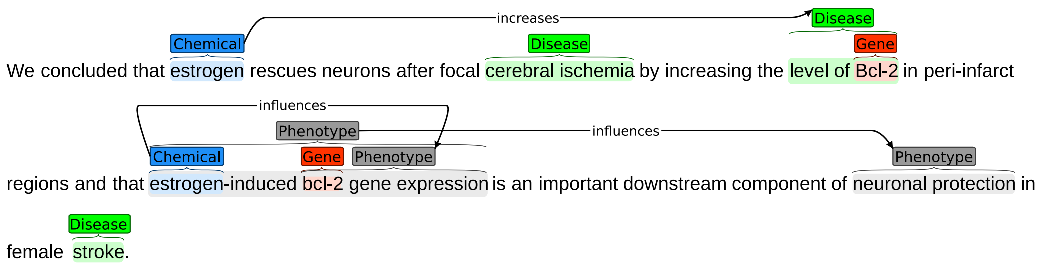
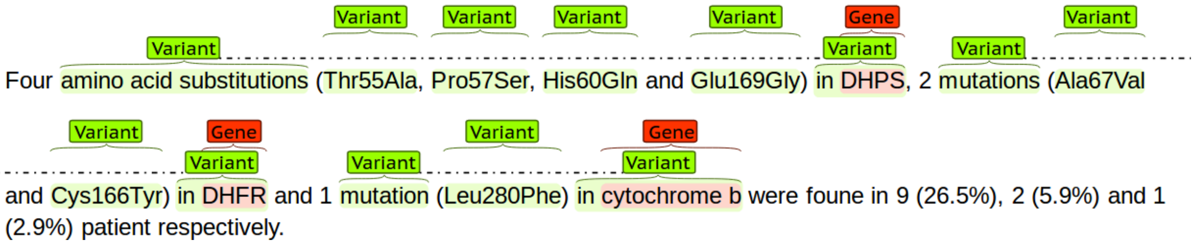

We concluded that estrogen rescues neurons after focal cerebral ischemia by increasing the level of Bcl-2 in peri-infarct regions and that estrogen-induced bcl-2 gene expression is an important downstream component of neuronal protection in female stroke.

Figure 4: First example

TODO: Check new relations in figure.

In this example (4),
“estrogen” is referred to as a Chemical, and not a body made substance that’s why it is annotated as such.
“Cerebral ischemia” is a disease. The word“cerebral” is annotated into it because it defines the disease in a strong way and gives sense to it.
“Bcl-2” is a protein, annotated as Gene as it as been decided for this corpus. It’s level (“level of Bcl-2”) is a Phenotype.
”Infarct” is a Disease.
“Estrogen-induced bcl-2 gene expression” is the expression of the bcl-2 gene that has been induced by estrogen. It is annotated as a phenotype.
“bcl-2 gene expression” is also a phenotype. It is the second part of an ‘nfluences” relation
“Neuronal protection” is a Phenotype.
“Stroke” is a Disease.

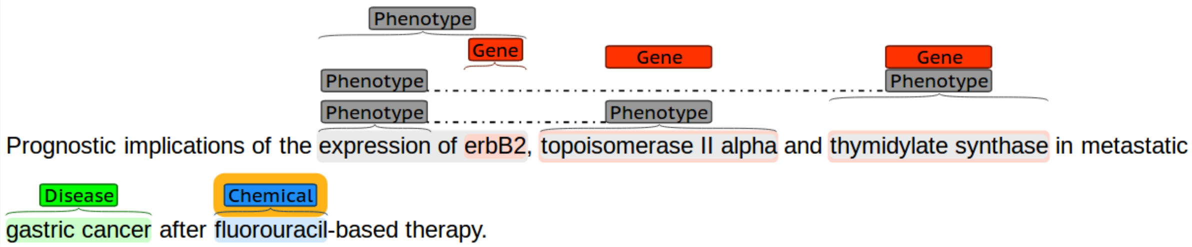
The sentence states a relation between two entities (“estrogen-induced bcl-2 gene expression” and “neuronal protection”). It reveals that the first is an important part of the second entity and thus, influences it, that’s why the relation ”Influences” is used.



Four amino acid substitutions (Thr55Ala, Pro57Ser, His60Gln and Glu169Gly) in DHPS, 2 mutations (Ala67Val and Cys166Tyr) in DHFR and 1 mutation (Leu280Phe) in cytochrome b were foune in 9 (26.5%), 2 (5.9%) and 1 (2.9%) patient respectively.

Figure 5: Second example

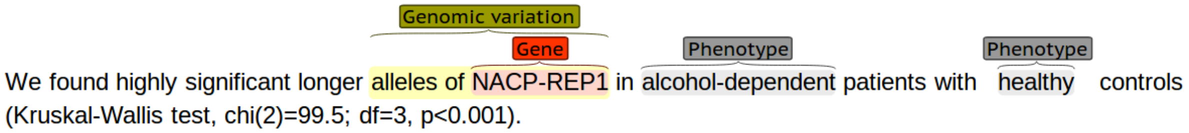
In this example (5), we can see a lot of "Variant" entities. They have two forms in this sentence.
The first is "amino acid substitutions ... in DHPS". This clearly states another form of the protein DHPS which refers to a genetic variation. That's why it is annotated as a Variant.
The second is "Thr55Ala" which is structured as : AminoAcid Number AminoAcid. It obviously is referring to a genetic variation too and is annotated as a Variant for this reason.
Finally, "DHPS" and "cytochrome b" are proteins annotated as "Gene". There is no relation in this sentence as it simply states results of a study.



Prognostic implications of the expression of erbB2, topoisomerase II alpha and thymidylate synthase in metastatic gastric cancer after fluorouracil-based therapy.

Figure 6: Third example

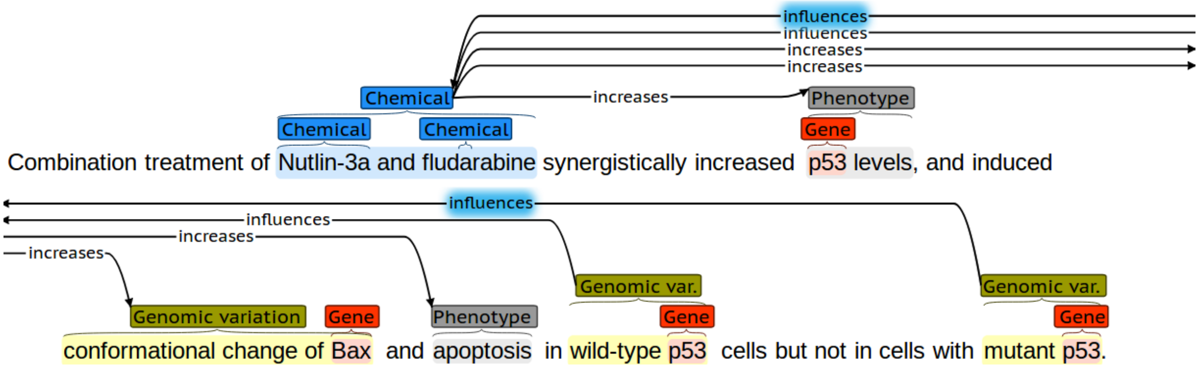
In this example (6), we can see three occurrences of the same kind of entities. These entities are Phenotype enclosing Gene.
First of all, we can see three proteins annotated as Gene.
Then, their expression is annotated as Phenotype.
Finally, the end of the sentence mentions "gastric cancer" which is a Disease and "fuorouracil" that is used as some a drug and will be annotated as a Chemical. No relationship in this sentence, even thought the sentence contains the "implications" word. The "prognostic" word make it too hypothetical to be annotated.



We found highly significant longer alleles of NACP-REP1 in alcohol-dependent patients with healthy controls (Kruskal-Wallis test, chi(2)=99.5; df=3, p<0.001).

Figure 7: Fourth example

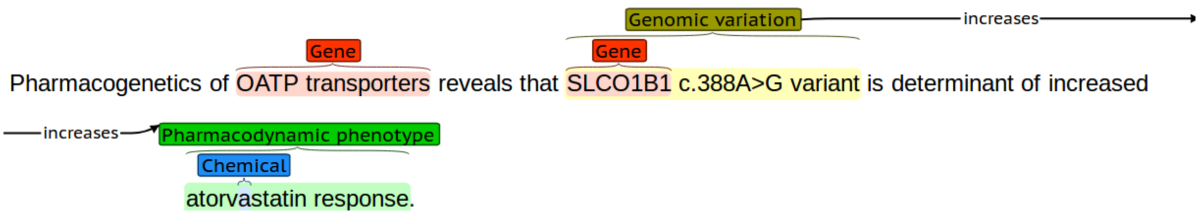
In this example (7), we can find a Gene ("NACP-REP1") and a Genomic variation that is attached to it ("alleles of NACP-REP1").
"Alcohol-dependent" is a Phenotype, and "Healthy" too.
In this sentence, the pre-annotation tool annotated "alcohol" as a Chemical, but as it is referred to as an element of the daily life, and not as a medical substance, it will not be annotated and the Chemical tag has been discarded. In this example (8), we can see two Chemical entities : "Nutlin-3a" and "fludarabine". The sentence states that it is used in a combined way to work synergistically. That is why they are annotated together as one Chemical.



Combination treatment of Nutlin-3a and fludarabine synergistically increased p53 levels, and induced conformational change of Bax and apoptosis in wild-type p53 cells but not in cells with mutant p53.

Figure 8: Fifth example

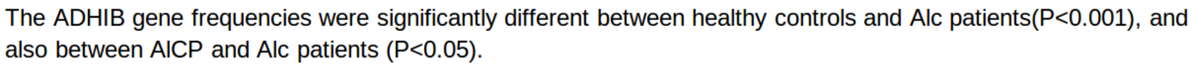
We can find the Gene "p53" in the Phenotype "p53 levels", the protein (Gene) "Bax" in the Genomic variation "conformational change of Bax", the Gene "p53" again in the Genomic variation "wild-type p53" and the Gene "p53" in the Genomic variation "mutant p53".
"Apoptosis" is a Phenotype.
There is a lot of relationships in this sentence.
The Chemical at the beginning of the sentence is said to increase (Increases) the two first Phenotype entities and the first Genomic variation entity. We have no real information about the way it influences the "wild-type p53" so the Influences relation is used.
Finally, it is explicit that the Chemical at the beginning of the sentence DOES NOT influence the "mutant p53" Genomic variation stated at the end of the sentence, that's why a negative (Neg) Influences relation is used between them.



Pharmacogenetics of OATP transporters reveals that SLCO1B1 c.388A>G variant is determinant of increased atorvastatin response.

Figure 9: Sixth example

In this example (9), we can see two Gene entities at the beginning of the sentence (”OATP” and ”OATP transporters”). They are proteins annotated as Gene.
”SLCO1B1” is a Gene stated in a Genomic variation entity (”SLCO1B1 c.388A>G variant”).
At the end of the sentence, we can find a Pharmacodynamic phenotype (”atorvastatin response”) with a Chemical entity attached to it (”atorvastatin”).
A relationship is clearly stated in this sentence (”is determinant of increased...”) so the increases relationship is used.



The ADHIB gene frequencies were significantly different between healthy controls and Alc patients(P<0.001), and also between AICP and Alc patients (P<0.05).

Figure 10: Seventh example

In this example (10), the sentence is not about pharmacogenomics. Thus, it has been decided to discard it and that is why no entity is stated in it.

References

[1] L. Campillos et al. Annotation scheme for the MERLOT French clinical corpus, 2017.

[2] P. Martnez M. Herrero-Zazo, I. Segura-Bedmar. Annotation guidelines for DDI corpus, 2015.

[3] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.

[4] Chih-Hsuan Wei, Bethany R. Harris, Donghui Li, Tanya Z. Berardini, Eva Huala, Hung-Yu Kao, and Zhiyong Lu. Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in pubmed abstracts. *Database(oxford)*, 18, 06 2012.