Annotation scheme: the Microbiota corpus

1 Introduction

Harnessing the potential of the latest information retrieval techniques and methods relies highly on annotated corpora. As we step in domains of speciality, gold standard corpora are fairly uncommon: it is the case in sub-domains of the biomedical area. A good deal of studies aimed to publish corpora focusing only on one or two concepts (entity type) along with one type of relation between the mentioned concepts (binary classification of the presence of relation between concepts). Therefore, we intend to introduce a corpus gathering different concepts strongly correlated with our study domain, being the human microbiome, along with the relations among them.

2 Data selection

For simplification purposes, we select full text articles provided by PubTator along with their annotations. These includes 6 concepts: Gene, Disease, CellLine, Species, Mutation and Chemical. The article selection process relies on the following query: ”gut microbiota OR gut microbiome OR intestinal microbiota OR intestinal microbiome”. Based on the hypothesis stating that a relation is potentially phrased in a sentence/paragraph only if at least two named entities are present in the same sentence/paragraph, we select articles based on a number of heuristics such as length of the paragraphs, number of named entities per label, and the number of unique mentions per label.

3 Annotation tool

For this task, Label studio is the selected tool for annotation. We integrate the provided annotation by PubTator in the mentioned concepts. The tool allows annotators to modify or delete PubTator annotations only if the impact on the rest of the text is important(for instance if an important interaction is expressed but one of the entities participating in the relation are not tagged).

4 Curation task

The aim of the annotation process is to label all relationships and possible interactions between the following entities: Species, Disease, Chemical, Gene, CellLine and Mutation. Each paragraph represents a unit of meaning, implying that the relation may be expressed anywhere within the unit (in the same sentence as the named entities, or across the other sentences). The unit may also be a single sentence. Relations may be found anywhere in the text unit, including inter-sentences. This implies that participating entities in a specific interaction may be expressed in different sentences within the same paragraph.

As mentioned previously, the annotation tool enables altering the annotated entities, including deletions, modifications and additions. This said, annotators may have access to different information resources only for clarification purposes. The annotations should not be based on external knowledge. The information, whether it concerns entities or interactions, should be explicitly cited in the text.

This said, the curation process consists of 2 major steps:

1. Revision of entities annotations of PubTator.

2. Annotation of the relationships.

5 Concept types

5.1 Definitions

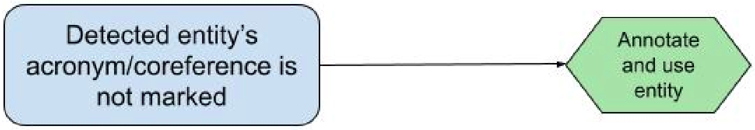
We introduce the selected concept types for this corpus. They represent the most relevant entities for our study of the human gut microbiome. The following table reveals the definitions, as well as the elements that should be considered as a concept for each one.

Concept type	Definition	Precision
Species	Group of living organisms consisting of similar individuals capable of exchanging genes or interbreeding	includes bacteria and protein
Chemical	Substance used in a chemical process or made by a chemical process.	includes substances used as medication or in the preparation of medication
Disease	Disorder of structure or function in a human, animal, or plant, especially one that has a known cause and a distinctive group of symptoms, signs, or anatomical changes.	includes symptoms and complications
Gene	Specific sequence of nucleotides in DNA or RNA that is located usually on a chromosome and that is the functional unit of inheritance	-
CellLine	Cell culture selected for uniformity from a cell population derived from a usually homogeneous tissue source (such as an organ)	include organs
Mutation	Change in the DNA sequence of an organism	-

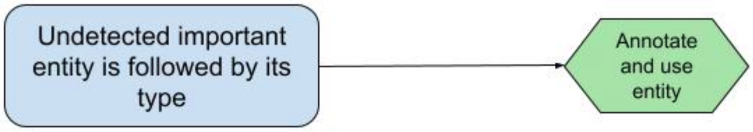
5.2 Guidelines and rules

The following cases describe situations you may encounter while annotating along with the expected behavior. It is important to note that any modification of named entities is necessary only when they are part of an important relation that has to be annotated.

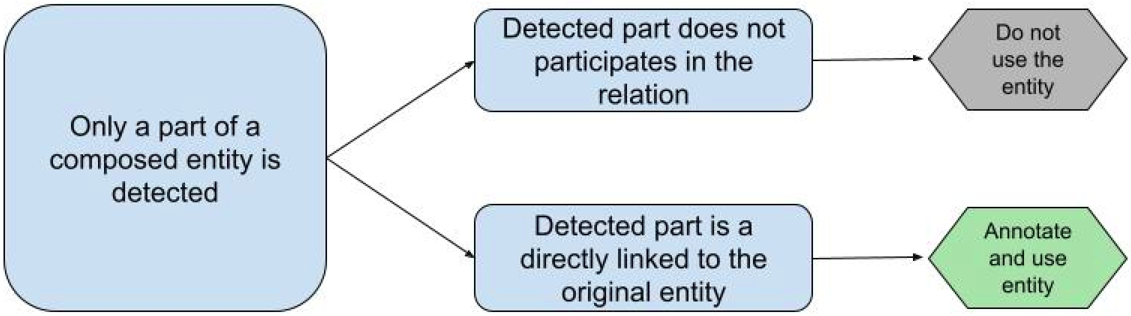
Case 1: A named entity is detected in the text, and is followed by its acronym that is not detected but used somewhere in the text where the relation is expressed.



Case 2: An entity is followed by its type in the text (e.g: alpha2macroglobulin gene), participates in a relation but is not detected.



Case 3: A part of a composed entity is detected, and its type is different from the one of the entire entity.


Only a part of a composed entity is detected
Detected part does not participates in the relation
Do not use the entity
Detected part is a directly linked to the original entity
Annotate and use entity

General rules :

- Annotate according to the context of the paragraph or sentence.
- Annotate bacteria as Species.
- "Gut" is annotated as Gene when followed by "microbiota": correct the annotation by annotating "gut microbiota" as Species.
- Viruses can be annotated as Species or Disease: The correct annotation depends on the context.
- Annotate chemical processes as Chemical (e.g median hepatic DNL)
- Annotate symptoms as Disease

6 Interactions types

6.1 Definitions

The interaction types between pairwise entities are partly derived from the UMLS Semantic Network. They describe interactions that occur between the following entities mentioned in the previous section.

Interacts_with

May occur between all entities.
Definition: Acts, functions, or operates together with. Used when the type of interaction is not specified.

Complicates

|all entity types|Complicates Disease

Definition: induces a complication, a medical problem that occurs during, or after a procedure, treatment or illness /Causes to become more complex.

Treats

|all entity types|Treats Disease

Definition: expresses a cure or a remedy for a symptom or disease.

Worsen

May occur between all entities

Definition: indicates a negative change in the aspect of an entity.

Affects

May occur between all entities.

Definition: brings about a direct effect on. It implies the influence and the impact of a new or existing state or entity on another one, including the following interactions: has a role in, contributes in, leads to, alters and modifies.

Example: ACC(Gene) converts acetyl-CoA(Chemical) to malonyl-CoA.(Gene Affects Chemical)

Causes

May occur between all entities.

Definition: Expresses a causal interaction, capturing a condition or a result. It implies that an entity has a direct impact or triggers a second entity or event.

Example: NAFL is a generic term that includes a series of liver diseases(Disease) with different injury severities and consequent fibrosis(Diseases).

Experiences

Species experiences Disease
Species experiences Chemical

Definition: induces that a specie is subjected to a condition or affected by a chemical serving as medication.

Example: patients(Species) with histologically proven NASH (Disease)(Patients Experiences NASH).

Improve

May occur between all entities.

Definition: indicates an improvement of a described aspect.

Increase/Decrease

May occur between all entities.
Definition: indicates an increase/decrease in a described aspect directly linked to an entity.
Example: he results showed that the administration of MK-4074(Chemical) for 1 month reduced liver TG(Chemical) by 36% in patients with hepatic steatosis(Chemical Decrease Chemical).

Start/Stop

May occur between all entities.
Definition: indicates an initiation/ending of a described aspect directly linked to an entity.
Example: The development of NAFL/NASH(Disease) is considered to initiate from simple steatosis(Disease) (Here, we annotate the interaction between steatosis and NAFL/NASH as START).

Negative_correlation

May occur between all entities.
Definition: expresses a negative association or link between the mentioned entities.

Reveals

May occur between all entities.
Definition: designates the discovery of the presence of an entity.

Prevents

Gene|Chemical|Species|Mutation|CellLine Prevents Disease.
Definition: describes an action taken to decrease the chance of getting a disease, a condition or symptom.

Physically_related_to

May occur between all entities.
Definition: expresses a relation based on a physical attribute or factor.
Example: ACC converts acetyl-CoA(Chemical) to malonyl-CoA(Chemical).

Used_for

Chemical used for |all entity types|.
Definition: designates the usage of chemical to affect physically an entity type.

Possible

May occur between all entities.
Definition: expresses a potential interaction between entities in need of confirmation.
Example: However, HCC(Disease) may also develop in the absence of cirrhosis(Disease)(the link between the two diseases is uncertain)

Presence

May occur between all entities.

Definition: indicates the co-presence of the entities.
Example: While ALD(Disease) is defined by the presence of hepatic steatosis(Disease) associated with significant alcohol consumption, NAFL is a generic term that includes a series of liver diseases.

Location_of

May occur between all entities.
Definition: indicates that the interaction is defined by the location.

Marker/Mechanism

Chemical is marker/mechanism of Disease.
Gene is marker/mechanism of Disease.
Definition: indicates that a gene or chemical is a biomarker of a disease or correlates with it. (e.g increased abundance in the brain of chemical X correlates with Alzheimer disease)

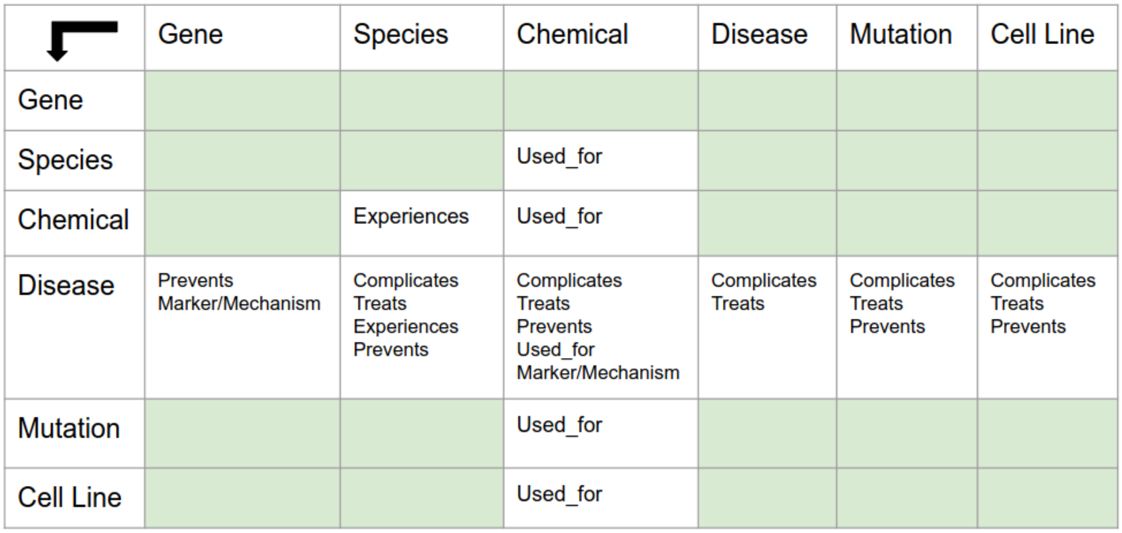
Part_of

May occur between all entities.
Definition: Specifies that an entity is involved physically in the being of another.
Example: NAFL(Disease) is a generic term that includes a series of liver diseases(Disease).

Associated_with

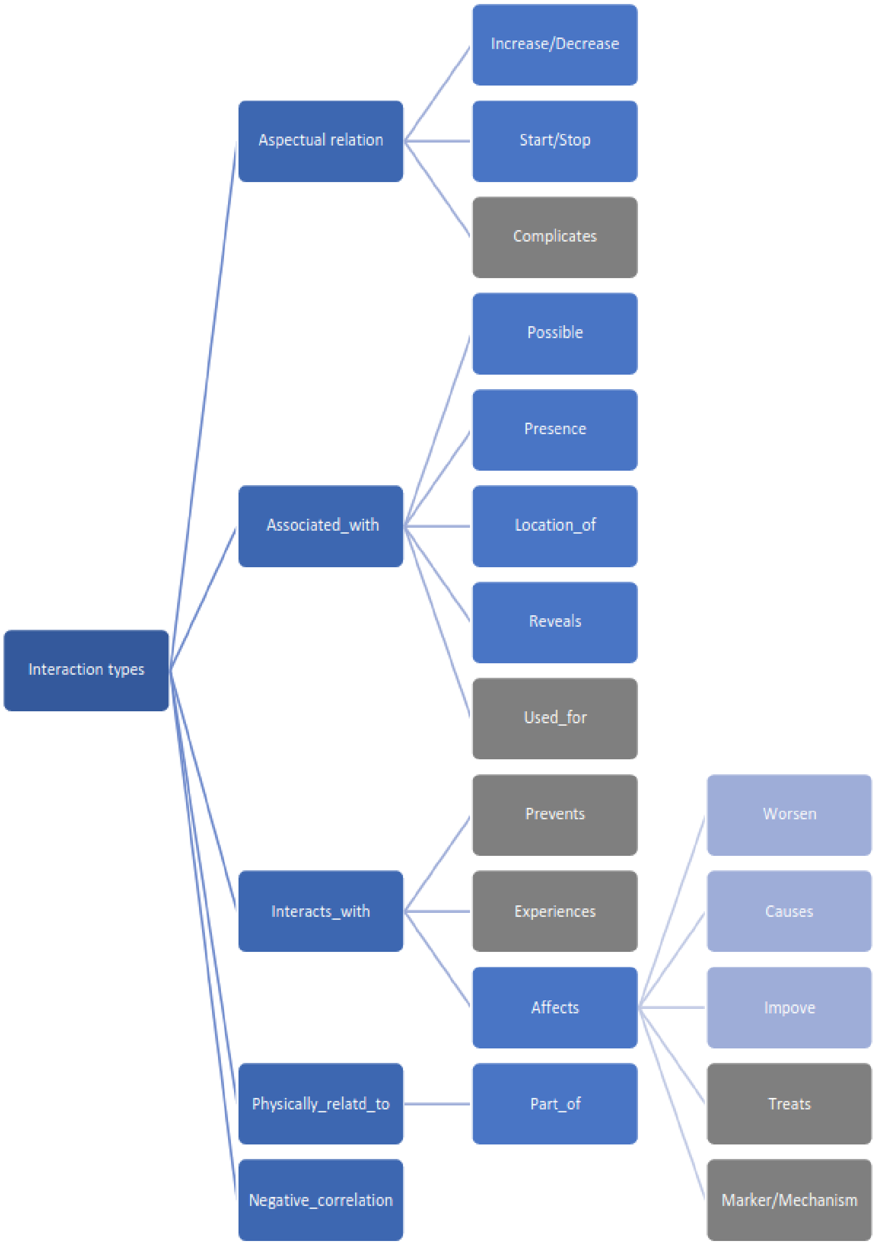
May occur between all entities.
Definition: designates the permanent/ direct link between two entities. Example: While ALD is defined by the presence of hepatic steatosis(Disease) associated with significant alcohol consumption(Disease), NAFL is a generic term that includes a series of liver diseases.

It is observable that most of relations may take place between all entities. In order to clearly visualize exceptions, we mark in the table below the relations occurring between specific entities.



⬐	Gene	Species	Chemical	Disease	Mutation	Cell Line
Gene						
Species			Used_for			
Chemical		Experiences	Used_for			
Disease	Prevents Marker/Mechanism	Complicates Treats Experiences Prevents	Complicates Treats Prevents Used_for Marker/Mechanism	Complicates Treats	Complicates Treats Prevents	Complicates Treats Prevents
Mutation			Used_for			
Cell Line			Used_for			

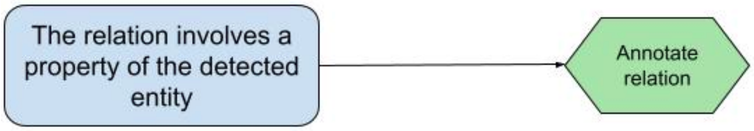
As mentioned before, the relations' annotation has to be as precise as possible. This induces that relations can be ordered from the least to the most specific. The hierarchical graph below maps the interaction types according to this claim. The grey boxes contain relations occuring between specific entities, while the rest of the boxes may appear between all types of entities.



6.2 Guidelines and rules

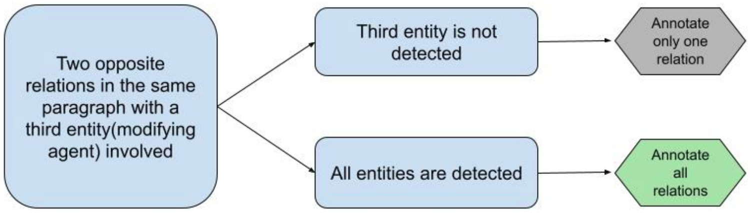
Case 1: The expressed interaction involves an attribute of the entity instead of the direct entity.

Example: The short chain fatty acid propionate (Chemical) stimulates GLP-1(Gene) and PYY(Gene) secretion via free fatty acid receptor 2 in rodents: The mentioned chemical stimulates the gene's secretion, which is the property of the mentioned gene.



Case 2: In the same paragraph, two opposite relations are expressed between the same entities. The opposition comes from the intervention of a third entity.

Example: The combined treatment with chemical 1 and 2 upregulates gene X, but not the single one.


Two opposite relations in the same paragraph with a third entity(modifying agent) involved
Third entity is not detected
Annotate only one relation
All entities are detected
Annotate all relations

The following rules are to consider when adding a relation between entities.

- Be as specific as possible when choosing relation labels.
- Annotate the relations expressed in the context where entities appear, not based on the general idea of the paragraph or external knowledge.
- Annotate inter-sentence relations.
- Annotate the relations that were just investigated in the text using the relation "possible".
- More than a single relation may occur between the same two entities.
- Prior knowledge may be used for confirmation.
- Annotate negative relations using the label "negative_correlation".
- Use "Interacts_with when the relation's nature is ambiguous.
- Do not annotate dosage (of chemicals).
- Do not annotate a statement of the absence of a relation as negative_correlation.