Annotation Guideline

(Version 1.0)

Abstract

This paper is a guideline that instructs annotators how to think and carry out tasks relating to the SNPPhenA corpus. After introducing the task description and the background, it explains how to annotate the corpus step by step.

Contents

1- Introduction

This document contains a description of the design principles underlying the SNPPhenA corpus and detailed information about the encodings, markup conventions and the linguistic annotation with which the corpus was enriched.

The SNPPhenA corpus was formed during a project that aimed to produce a software package for automatically determining the existence of an association between various polymorphisms and relevant phenotypes or traits that are presented in academic articles. This corpus can mainly be used for the task of extracting biomedical relations and the degree of associations from scientific texts with accuracy and confidence. In general, the task of biomedical relation extraction is focusing on biomedical entities such as proteins, drugs, and genes in biology related article texts and trying to find binary or complex relations between them.

2- Structure

The SNPPhenA corpus consists of 360 documents in xml format and encoded in utf-8; each of which is drawn from the abstract sections of papers published in journals and conferences related to the subject of life sciences. These documents have been annotated according to the conventions presented in this reference.

2-1- Markup conventions

Building blocks and valid tags for the documents can be found on the document type definition file named SNPPhenA.dtd, which is located in the corpus folder.

The SNPPhenA is delivered through UTF-8 encoding. Almost all characters in the corpus are represented directly by the appropriate Unicode character. Some exceptions are as follows:

- the ampersand (&), which is represented by the special string &
- the double quotation mark, which is represented by the special string "
- the arithmetic less-than sign, which always appears as <
- the arithmetic less-than-or-equal sign, which always appear as ≤
- the arithmetic greater-than sign, which always appear as >
- the arithmetic greater-than-or-equal sign, which always appear as ≥

In the SNPPhenA corpus, the names of tags have been written in lowercase; while the attribute names have been put into uppercase. Each document in the corpus has a unique id in the whole corpus, starting from “1000”. IDs of critical sentences are a composition of the

document id followed by a local id, which is numbered from 0, such as “1047_0” in the above example. IDs for the other blocks are defined to be unique only on the containing document and so, start from 0.

The START and END attributes for each tag indicate the index of the beginning and the end of the referring string in the original text.

2-2- An example

Here is a complete example of an annotated document. The example begins with the start tag for an `<abstract>` element, which bears an abstract id attribute - the value of which is 1047 - and a text attribute, representing the abstract text. The start tag is followed by two `<sentence>` elements, which provides the critical sentences from the original source text. Sentence elements in turn are followed by one or more `<snp>`, `<phenotype>`, `<modality_marker>`, `<negation_scope>` and `<pair>` elements. Each of these tags is described in the next few sections.

```
<abstract ABSTRACTID="1047" TEXT="OBJECTIVE: There is compelling evidence that the plasma apolipoprotein E (APOE) concentration, in addition to the APOE ε2/ε3/ε4 genotype, influences plasma lipoprotein levels, but the functional genetic variants influencing the plasma APOE concentration have not been identified. APPROACH AND RESULTS: Genome-wide association studies in 2 cohorts of healthy, middle-aged subjects identified the APOE locus as the only genetic locus showing robust associations with the plasma APOE concentration. Fine-mapping of the APOE locus confirmed that the rs7412 ε2-allele is the primary genetic variant responsible for the relationship with plasma APOE concentration. Further mapping of the APOE locus uncovered that rs769446 (-427T/C) in the APOE promoter is independently associated with the plasma APOE concentration. Expression studies in 199 human liver samples demonstrated that the rs769446 C-allele is associated with increased APOE mRNA levels (P=0.015). Transient transfection studies and electrophoretic mobility shift assays in human hepatoma HepG2 cells corroborated the role of rs769446 in transcriptional regulation of APOE. However, no relationships were found between rs769446 genotype and plasma lipoprotein levels in 2 cohorts (n=1648 and n=1039) of healthy middle-aged carriers of the APOE ε3/ε3 genotype. CONCLUSIONS: rs769446 is a functional polymorphism involved in the regulation of the plasma APOE concentration.">
    <sentence ID="1047_0">
        <snp ID="0" START="1175" END="1183" TEXT="rs769446"/>
        <phenotype ID="0" START="1197" END="1315" TEXT ="plasma lipoprotein levels"/>
        <modality_marker  START="1161" END="1166" TEXT ="found"/>
        <negation_scope  START="1139" END="1315">
            <negation_cue START="1139" END="1141" TEXT ="no"/>
        </negation_scope>
        <pair PAIRID="0" PHENOTYPEID="0" SNPID="0" ASSOCIATION="negative" CONFIDENCE=""/>
    </sentence>
```

```
        <sentence ID="1047_1">
                <snp ID="1" START="1329" END="1337" TEXT ="rs769446"/>
                <phenotype ID="1" START="1401" END="1427" TEXT ="plasma APOE concentration"/>
                <modality_marker  START="1316" END="1327" TEXT ="CONCLUSIONS"/>
                <pair PAIRID="1" PHENOTYPEID="1" SNPID="1" ASSOCIATION="positive"
                CONFIDENCE="medium"/>
        </sentence>
</abstract>
```

3- Tags

A basic element in each document is `<sentence>`. Any sentence in the original text that contain at least one SNP and phenotype entity is considered to be critical sentence and is annotated with the appropriate tags. These tags include entity tags, relationship tags and features tags. Each of them is explained below:

3-1- Entities

Two main classes of the entities considered in this corpus consist of SNPs and phenotypes. From the scientific point of view, SNPs are a variation in a single nucleotide that occurs at a specific position in the genome. For the annotation task, all known SNP names as well as any mention of them in the text, which refers to their famous gene symbols, are selected as the `<snp>` entity. These names mostly come from the open-access databases including [1]: SNP500Cancer, SNPedia [2], and pharmGKB [3].

In the following example, “rs429358” and “rs7412” are the name of two SNPs that are clearly expressed in the text. For simplicity, tags for SNPs and phenotypes have been embedded in the original text.

> Example 1
>
> “Apolipoprotein E (APOE) functional haplotypes determined by `<snp>`rs429358`</snp>` and `<snp>`rs7412`</snp>` SNPs have been extensively studied and found to be one of the most consistent association in human `<phenotype>`longevity`</phenotype>` studies.”

Different types of human characteristics should be tagged as `<phenotype>`, which includes a wide range of unusual circumstances from traits to diseases. Indeed, a phenotype is the appearance of an organism in terms of traits such as its morphology, development,

physiological properties, behavior, and products of that behavior [4]. Two more complete related databases were chosen for this task: a list of Comparative Toxicogenomics Database (CTD) for disease names [5], and the phenotype ontology that was prepared for the blast project [6]. The collected list of phenotypes includes 65,530 phenotype names with more than twelve thousand disease names and their synonyms. In example 1, “longevity” is a phenotype. The following table gives an example for the designed entities.

Entity type	Description	Example
Phenotype	Name dedicated to each abnormality or feature.	Coronary heart disease
SNP	rs plus number (rsID) and other corresponding historical numbers dedicated to each polymorphism, which ensure a probability of association with phenotypes.	rs499818, A1450G

3-2- Relationship

The main tag under the `<sentence>` is the `<pair>` tag. This tag represents the biomedical relation between a SNP and phenotype pair, which has been annotated with the appropriate tags in that sentence. Attributes defined for the `<pair>` tag includes an ID for uniqueness, the referred phenotype’s ID and the SNP’s ID, the association, and the strength of that association (degree of confidence) between the entities.

An association indicates the existence or lack of existence of a correlation between the appointed SNP and the relevant traits. Each SNP-phenotype pair falls into one of the following categories: 1) positive, 2) negative, and 3) neutral. If the critical sentence expresses an association between SNP and phenotype, in terms of a cause-effect relation between SNP and phenotype with a probability greater than zero, this pair of entities has a “positive” association. On the other hand, a “negative” association occurs when the SNP-phenotype pair evidently lacks any association. Additionally those pairs in the sentence which are not remarked upon due to an association or lack of association get a “neutral” value for the ASSOCIATION attribute.

Another attribute for the tag `<pair>` is CONFIDENCE, which is the greatest advantage of the SNPPhenA corpus. This attribute shows the degree of the association that is described in the ASSOCIATION attribute. When the association is positive, three levels of confidence are defined; “weak”, “moderate”, and “strong”. In general, different authors write the same fact about

the entities using different linguistic forms and with different levels of confidence. To annotate the confidence level, annotators should note the real value of the degree of the association; i.e. if it is presented in terms of the p-value in the sentence or in the paragraph. About 20% of sentences with positive pairs have this remark. If this matter is not apparent in the sentence, the tone of the writer and key words like modality markers and negation words should be considered and annotators therefore should determine the confidence level based on them. These cues are referred to as features and are presented at the next section.

If the association between entities is “neutral”, then the CONFIDENCE value is “zero” because the degree of the association is zero. In “negative” cases, there is no association and thus no strength in the association. For simplicity, the CONFIDENCE value is presented with “-” in the corpus.

> Example 2
> <snp>Apolipoprotein E (ApoE) genotype</snp> has been associated with <phenotype>systemic inflammation</phenotype> and athero-thrombosis however the association with <phenotype>abdominal aortic aneurysm</phenotype> (AAA) has not been previously examined.”

In the above example, it is understandable that there is an association between “ApoE” and “systemic inflammation”; so this pair has a positive association. In contrast, the pair “ApoE” and “abdominal aortic aneurysm” belongs to the neutral category. Since the author says this association has not been examined, this pair cannot therefore be labeled as positive or negative.

> Example 3
> “The genetic factors studied were not associated with cognitive status in PD patients. Only age and Hcy plasma levels were found to be independent risk factors predisposing individuals to PD dementia. However, <snp>COMT: rs4680: A>G </snp> and rs4633: C>T polymorphisms were found to significantly affect <phenotype>PD</phenotype> risk, and the <snp>MTHFR 677C>T</snp> polymorphism helped determine <phenotype>plasma Hcy concentrations</phenotype>.”

In example 3, which is drawn from the conclusion section, there are two SNPs (“COMT: rs4680: A>G” and “MTHFR 677C>T”) and also two phenotypes (“Parkinson's disease (PD)” and “plasma Hcy concentrations”). The result of the analysis has shown that the first

SNP affected PD; while the second one affected the next phenotype. As a result, these two SNP-phenotype pairs have a positive association. In contrast, an opposite SNP-phenotype combination, e.g. “COMT”- “plasma Hcy concentrations” and “MTHFR”-“PD”, have a negative association. Meanwhile, two positive pairs should be tagged with a degree of association. Relating to the first one, the findings showed a strong association as a result of the phrase “was found to significantly affect”. However, the second positive pair had a weak association due to the phrase “helped determine”.

3-3- Features

Each critical sentence should be enriched using some informative tags such as <modality_marker> and <negation_scope> and <negation_cue>. These kinds of annotations seem to be effective for determining the degree of the association between the entities and the categories of the pairs. As a result, this information may be used alongside appropriate machine learning algorithms for extraction.

3-3-1- Modality

Modality markers are those words that qualify the opinion and attitude of the speaker or writer. The writer of an academic paper can make judgments about the truth of a proposition or state a fuzzy proposition which is true in part on some occasion by utilizing modality marker words such as “may”, “could”, “possibly”, “almost”, “indicate” and “found”.

For example consider the following sentence. In this sentence, the word that indicates the existence or lack of existence of an association between “rs769446” and “plasma lipoprotein levels” is “found”.

Example 4:

“No relationships were <modality_marker>found</modality_marker> between <snp>rs769446</snp> genotype and <phenotype>plasma lipoprotein levels</phenotype> in 2 cohorts (n=1648 and n=1039) of healthy middle-aged carriers of the APOE ε3/ε3 genotype.”

As another example, the following sentence represents an association between “APOA5-1131C” and “MI”, as determined by the phrase “strongly affects”.

Example 5:

"The <snp>APOA5-1131C</snp> allele, associated with higher fasting triglyceride levels, <modality_marker>strongly</modality_marker> affects the risk for early-onset <phenotype>MI</phenotype>, even after adjusting for triglycerides.

Each critical sentence that has modal words should be annotated by <modality_marker> tags. The employed modality markers have been obtained from the list that is provided in [7]. This list is an extension of the list presented in [8] for biomedical domains.

3-3-2- Negation

Critical sentences containing any kind of negation are tagged with a <negation_cue> and a <negation_scope>. Negation is understood as the implication of the non-existence of something. However, the presence of a negative word does not imply that the pairs in these sentences should be annotated as "negative" cases. The annotators must pay attention to the sentences that include negative words.

The scope of a <negation_cue> tag starts directly at the beginning of the key word, and ends at the end of the key word. Lists of these key words are available to the annotators. Negation cues can occurs in different morphological types, such as verbs like (lack), adverbs (not), adjectives (absent), determiners (no), nouns (absence), conjunctions (neither), and prepositions (without) [9].

The scope of a <negation_scope> tag can extend to the whole sentence containing <negation_cue> tags, or to certain phrases. Different negation cues in different structures, such as active and passive sentences, have different scopes of negation. The instruction for annotating negation scopes is adopted from the rules given in the work of Morante [9].

4- Annotation Process

One simple method for annotating a document is to read the document from start to end and mark the annotation in the order in which they appear. However, this does not result in the most accurate corpus. In order to gain consistent annotations, it is necessary to have a methodology, according to which all annotators think and do the same thing. Therefore, all annotators were asked to perform these steps in order:

1- *Read the whole document;* Reading the whole document without thinking about entities or relationships is necessary to get an understanding of it.

2- *Mark the entities*; If the tags of entities are incorrect, or some entities have no tag, annotators should edit the tags. (It should be noted that the entities have been already tagged using automatic tools.)
3- *Mark features;* If tags of features are incorrect, or some features have no tag, annotators should edit the tags.
4- *Find critical sentences;* If a sentence or other close sentences which speak about the same entities contain both a SNP and a phenotype, mark these sentences as critical sentences.
5- *Find the relations;* Considering entities and features in a critical sentence, annotators should focus on relations between each pair of entities to determine the existence of associations and the confidence level of the author. Annotators should follow the instructions while annotating each of the tags described before.
6- *Look again*; Reviewers are asked to be certain that nothing is missed. In particular, annotators should count the number of pairs that they tagged. If there is x SNP and y phenotype in a critical sentence, they should annotate $x \times y$ pairs.
7- *Record any question, or ambiguous situation;* If reviewers have any questions that need to be clarified, they should record them.

After the annotators completed their task, a software program was hired to find inconsistencies and missing items. Furthermore, inter-agreement analysis has been performed to measure the quality of the annotations.

5- Frequently Asked Questions

In this section, the most frequently asked question of the annotators are presented.

5-1- How do we determine the confidence level if the degree of the association between entities are presented in terms of the p-value?

If the p-value is lower than or equal to 0.001 ($p \leq 0.001$), CONFIDENCE is “strong”. If the p-value is greater than 0.001 and less than 0.01 ($0.01 < p < 0.001$), CONFIDENCE is “moderate”. In the case of the p-value being greater than or equal to 0.01 ($p \geq 0.01$), CONFIDENCE is “weak”. As an example, consider the following sentences. The degree of the association between narcolepsy and rs5770917 is “weak” because the p-value is 0.02.

Example 6:

"<snp>rs5770917</snp>, a SNP located between CPT1B and CHKB, was associated with <phenotype>narcolepsy</phenotype> in Japanese (<snp>rs5770917</snp>[C], odds ratio (OR) = 1.79, combined P = 4.4 x 10(-7)) and other ancestry groups (OR = 1.40, P = 0.02)."

5-2- What is the scope of the negation cues?

Some of the more important negation cues and their function and scope are as follows:

- "No" is the most frequent negation cue in clinical reports. "No" occupies the first position of a nominal sentence. The full noun phrase in which "no" is a determiner is under the scope of the negation. If "no" modifies a noun, it scopes over the noun phrase and if it modifies an adjective, it scopes over the adjectival phrase.

  "<negation_scope> <negation_cue>No</negation_cue> association between COMT and smoking behavior was observed</negation_scope>"

- "Not" is always a negation cue. If it modifies a verb, it scopes over the verb phrase in active sentences and over the clause in passive sentences. If it modifies other phrases, it scopes over the phrase.

  "<negation_scope> <negation_cue>not</negation_cue> confirm the association of CPT1B/CHKB (rs5770917) in the Chinese population"

- In active sentences, negation cues, such as "cannot", "could not", "didn't", and "exclude" scope over the object of the main verb, and also over the subject in passive sentences.

  "COMT <negation_scope><negation_cue>didn't</negation_cue> affect CPP </negation_scope>"

- "Neither … nor" scopes over the full clause, if it coordinates copulative clauses or clauses in a passive form.

  "<negation_scope> <negation_cue>neither</negation_cue> rs11196218 <negation_cue>nor</negation_cue> rs290487 showed a significant association </negation_scope>"

- The determiner "neither" always acts as a negation cue, which scopes over the full clause.
- The noun "absence" is always a negation cue whose scope is the prepositional phrase "of that", which is required by "absence."
- The adjective "absent" scopes over the noun phrase it modified, or over the copulative clause if it participated in it.

5-3- How can the SNPPhenA corpus be extended?

This corpus can be extended qualitatively by using the complementary annotation tags, which may be useful for determining the degree of association and quantitatively by adding supplementary articles.

Acknowledgement

The author would like to acknowledge Dr. Mariana Neves (University of Potsdam) and Dr. MT Pilehvar (Cambridge University) for their useful comments.

References

[1] Bernice R Packer et al., "SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D617--D621, 2006.

[2] Michael Cariaso and Greg Lennon, "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis," *Nucleic acids research*, vol. 40, no. D1, pp. D1308--D1312, 2012.

[3] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, and et. al., "PharmGKB: the pharmacogenetics knowledge base," *Nucleic acids research*, vol. 30, no. 1, pp. 163-165, 2002.

[4] Elizabeth Martin and Robert Hine, *A Dictionary of Biology, 6 ed.*: Oxford University Press, 2014.

[5] Allan Peter Davis et al., "Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical--gene--disease networks," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D786--D792, 2009.

[6] (2015) Basic Local Alignment Search Tool (BLAST). [Online]. https://blast.ncbi.nlm.nih.gov/Blast.cgi

[7] Paul Thompson, Giulia Venturi, John McNaught, Simonetta Montemagni, and Sophia Ananiadou, "Categorising modality in biomedical texts," in *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, 2008, pp. 27-34.

[8] Ken Hyland, "Talking to the Academy: Forms of Hedging in Science Research Articles.," *Written Communication*, vol. 3, no. 2, 1996.

[9] Roser Morante, "Descriptive analysis of negation cues in biomedical texts," in *LREC*, 2010.

[10] Michael Seringhaus and Mark Gerstein, "Manually structured digital abstracts: A scaffold for automatic text mining," *FEBS letters*, vol. 582, no. 8, p. 1170, 2008.

[11] Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J Khoury, "A navigator for human genome epidemiology," *Nature genetics*, vol. 40, no. 2, pp. 124-125, 2008.