Importance of negations and experimental qualifiers in biomedical literature

Martin Krallinger
Struct. Biol. and Biocomp. Prog.
Spanish National Cancer Center, Madrid, Spain.
mkrallinger@cnio.es

Abstract

A general characteristic of most biomedical disciplines is their primarily experimental character. Discoveries are obtained through molecular biology and biochemical techniques that allow understanding of biological processes at the molecular level. To qualify biological events, it is of practical significance to detect specific types of negations that can imply either that a given event is not observed under specific conditions or even the opposite, that a given event is true by altering the bio-entities studied (e.g. introducing specific modifications like mutations). Of special interest is also to determine if a detected assertion is linked to experimental support provided by the authors. Finding experimental qualifier cues and detecting experimental technique mentions is of great interest to the biological community in general and particularly for annotation databases. A short overview of different types of negations and biological qualifiers of practical relevance will be provided.

1 Biological Annotations

In line with the rapid accumulation of biological literature and the growing number of large-scale experiments in biomedicine, it is becoming more important to capture essential facts contained in the literature and storing them in form of biological annotations. Such annotations usually consist in structured database records, where biological entities of relevance, like genes or proteins are associated to controlled vocabularies that are useful to describe the most relevant aspects of these entities (their function, localization, processes or pathways they participate in or implications in diseases). Also specific types of relations between bio-entities (e.g. physical or regulatory interactions) are manually extracted from the literature. For biological interpretation and to determine the reliability of annotations it is crucial to capture both negative annotations, whether a given relation has been studied experimentally and does not occur, as well as to determine the experimental method used to study the bio-entity of interest. For instance, the value of in vitro generated results, or those obtained by large-scale experiments have a different significance compared to those generated in vivo. The most relevant biological annotations contained in databases and constructed manually by expert curators are linked to experimental qualifiers. Such experimental qualifiers can range from simple method terms to more sophisticated ontologies or hierarchical terminologies. Experimental qualifiers used to annotate biological entities are for instance provided by the Proteomics Standards Initiative Molecular Interaction (PSI-MI) ontology, (Orchard S, Kerrien S., 2010) the Evidence Codes of Gene Ontology (GO) (Rogers MF, Ben-Hur A, 2010) or the Open REGulatory ANNOtation (ORegAnno) database Evidence Types.

2 Importance of Negations in Biomedicine

There is an increasing interest to extract from the literature negative associations. For instance, one of the most popular biological annotation efforts, Gene Ontology Annotation (GOA), also supports the annotation of $'NOT'$ relations (association.is_not) to be able to represent these types of relations in their annotation data. In GO, such relations are labeled using $'NOT'$ in the qualifier column for a particular annotation. This negation qualifier is applied to provide an explicit note that the bio-entity is not associated with a given GO term. This is important when a GO term might otherwise be expected to apply to a bio-entity, but an experiment proves otherwise. Negative asso-

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 46–49,
Uppsala, July 2010.

ciations are also used when a cited reference explicitly states a negation event, e.g. in the form of: bio-entity X is not found in the location Y. In addition to annotation efforts there are a range of scenarios where extraction of negative events are of practical importance, these are described in the following subsections.

2.1 Negations and Negative Controls

A common setting in experimental biology is to use controls to avoid alternative explanations of results and to minimize experimental artifacts. Negative controls corroborate that the experimental outcome is not due to some sort of unrelated effect; it serves to minimize false positives and can serve as a background observation. The underlying assumption of negative controls is that one assumes in advance that the result should be negative, i.e. no significant effect should be obtained. Such negative controls are mainly expressed in the literature using negations. For instance in case of protein-protein interaction experiments, a negative control could be to demonstrate that a signal is only obtained when the two interactor proteins are present, and not when the label (tag-protein) alone is given to each interactor individually. To illustrate this aspect consider the example sentences provided below:

- *Our results show that, when AGG1 is present in the matrix, it shows a strong ability to bind 35S-labeled AGB1, whereas GST alone is not able to bind any detectable AGB1.*
- *GST alone did not interact with FKHR even in the presence of E2 (Fig. 2B, lane 5), indicating the specific interaction between ER and FKHR.*
- *35S-labeled in vitrotranslated FBXO11 bound to immobilized GST-p53 (lane 3) but not GST alone (lane 2).*
- *PKC bound to GST-RINCK1 (lane 2) but not to GST alone (lane 1), revealing that PKC binds to RINCK directly.*

In those example cases, GST (alone) would represent the negative control. Only in presence of the interactor proteins a signal should be observed, if GST alone is present the assumption is that no signal should be obtained. Negative controls are crucial for interpretation of the actual experimental outcome.

2.2 Negative associations in medical and population genetics

A considerable effort is being made to detect genes and mutations in genes that have implications in the susceptibility of complex disorders. Naturally occurring variations in the sequence of genes, often called polymorphisms might have a deleterious, protective or no associations at all to a pathologic condition. Not only to capture deleterious and protective mutations, but also those that do not have any effect is important to aid in the interpretation of mutations observed in patients. This is especially true taking into account the increasing use of molecular screening technologies and personalized medicine in the clinical domain. Example cases of negative associations between genes and mutations to disease conditions derived from PubMed abstracts can be seen below:

- *CC16 gene may be not a susceptibility gene of asthmatic patients of Han population in southwest China.*
- *The FZD3 gene might not play a role in conferring susceptibility to major psychosis in our sample.*
- *Apolipoprotein E gene polymorphism is not a strong risk factor for diabetic nephropathy and retinopathy in Type I diabetes: case-control study.*
- *In view of this evidence, it is likely that the SIGMAR1 gene does not confer susceptibility to schizophrenia.*
- *Thus, this SNP in the PGIS gene is not associated with EH.*
- *The gene encoding GABBR1 is not associated with childhood absence epilepsy in the Chinese Han population.*
- *We did not find an association between OCD, family history for OCD, and the COMT gene polymorphism.*

Such negative associations can be useful for the interpretation of relevance of genes for certain conditions, enabling filtering un-relevant genes and improving target selection for more detailed molecular examinations.

2.3 Toxicology and negations

A simplified view of toxicology experiments is to distinguish, given the administration of different amounts of a specific compound or drug (e.g. low, medium and high dosage) during predefined time spans, between toxic and non-toxic effects. Such effects can be examined in animal models like rats or mini-pigs by examining a series of aspects, such as hematological parameters, organ histological properties (tissue alterations and size of organs), biochemical parameters, and changes in food/water consumption or fertility. Usually animals to which specific amounts of the compound has been administered are compared to control cases. Here it is important to determine also three kinds of negative associations: (1) under which conditions a given parameter or tissue has not been negatively affected (save dosage, non-toxic), (2) which compound did not show the desired beneficial effect (e.g. was not effective in treating the pathologic condition) and (3) under which administration conditions a compound was not save. Example sentences illustrating these negative associations are:

- *Morphological evaluation showed that 1-BP did not cause morphological changes in seminiferous epithelium, but 2-BP treatment resulted in the disappearance of spermatogonia, atrophy of the seminiferous tubules and degeneration of germ cells..*
- *This is an indication that the extracts may not be completely safe in male rats when continuously administered for 14days.*
- *Histopathologic analysis of the vital organs revealed no significant lesions in the brain, liver, kidney, heart, spleen, ovary, and testis.*
- *The extract did not produce any significant (P>0.05) changes in the mean concentrations of urea, creatinine, Na+, K+, and Cl- ions of rats in the extract treated groups compared to that of control.*

2.4 Experimentally altered bio-entities and negations

In order to characterize certain biological associations, it is a common practice to alter the bio-entity of interest, with the assumption that a given observation should change upon alteration. This is the case of mutations or deletions experimentally introduced to gene or protein sequences, with the underlying assumption that the mutated or truncated protein/gene should loose it ability to bind or regulate another bio-entity, or even be non-functional. Such mutations are useful to pin down the actual biologically relevant functional parts of bio-entities, which are usually of great therapeutic importance (as target sites to inhibit certain bio-entities or interactions). Such cases can be seen in the example sentences provided below:

- *Accordingly, this p73 N-terminal deletion was unable to activate transcription or to induce apoptosis.*
- *The G62D mutant did not bind AMP at all.*
- *The resulting mutant SOS3 protein was not able to interact with the SOS2 protein kinase and was less capable of activating it.*
- *MYB4 did not localize to the nucleus in the sad2 mutant, suggesting that SAD2 is required for MYB4 nuclear trafficking.*

In these example cases, altered bio-entities did not display the biological function of their wild type (unaltered) counterparts.

3 Experimental qualifiers

Biological annotation efforts are primarily concerned about experimentally confirmed events. Despite the importance of experimental qualifiers, only limited effort has been made to construct comprehensive resources to retrieve assertions that have experimental support and to construct useful lexical resources and thesauri of experimental evidence techniques. To detect novel protein interactions that have been experimentally characterized in the biomedical literature was one of the tasks posed in the BioCreative challenge, a community effort to assess text-mining tools developed for the biomedical domain (Krallinger M, et al, 2008). Also some systems to detect technical term mentions have been developed such as Termine. A range of recurrent cues relevant for experimental qualifiers can be observed in the literature, some of the most relevant ones are summarized in the table 1.

Using such experimental evidence cues together with linguistic patterns and NLP techniques it is feasible to determine whether a given event described in the literature has some sort of experi-

Cue	Pattern	PMID
reveal	METHOD revealed that EVENT	12506203
show	METHOD showed that EVENT	17189287
demonstrate	METHOD demonstrated that EVENT	18466309
study	EVENT was studied by METHOD	15147239
identify	EVENT identified in METHOD	10905349
prove	EVENT proved by METHOD	16354655
analyze	EVENT analyzed by METHOD	9477575
determine	EVENT determined by METHOD	12006647
confirm	EVENT confirmed using METHOD	10788494
obtain	EVENT obtained by METHOD	16582012
support	EVENT supported by METHOD	18156215
corroborate	EVENT corroborated using METHOD	15757661
validate	EVENT validated by METHOD	17287294
verify	EVENT verified by METHOD	18296724
detect	EVENT detected with METHOD	14581623
discover	EVENT discovered by METHOD	11251078
observe	EVENT observed using METHOD	16778013
test	EVENTwas tested using METHOD	14646219

Table 1: Experimental evidence cue terms.

mental qualifier associated to it. The simplest patterns of this sort would be for instance:

- *METHOD cue (a|that|novel|the|this)*
- *METHOD cue that*
- *as cue by METHOD*
- *was cue by METHOD*
- *cue (in|by|here by|using|via|with) METHOD*

Applying such patterns can be useful to construct automatically an experimental technique dictionary that can be handcrafted to enrich existing evidential qualifier resources. Nevertheless, linking automatically extracted experiment terms to controlled vocabularies used for annotation in biology is still a challenging task that need more manually labeled textual data. Some example sentences illustrating the usefulness of experimental evidence cues can be seen below:

- *Gel-shift and co-immunoprecipitation assays have revealed that GT-1 can interact with and stabilize the TFIIA-TBP-TATA complex.*
- *By yeast two-hybrid assays, we demonstrate an interaction of APC2 with two other APC/C subunits.*
- *The specificity of interaction of VIRP1 with viroid RNA was studied by different methodologies, which included Northwestern blotting, plaque lift, and electrophoretic mobility shift assays.*
- *A complex containing Mus81p and Rad54p was identified in immunoprecipitation experiments.*
- *In addition, we proved by affinity chromatography that NaTrxh specifically interacts with S-RNase.*

Acknowledgments

I would like to thank Yasmin Alam-Farugue (GOA team at EBI) for useful information on the annotation of negative associations in GOA and Roser Morante for important feedback and suggestions on this topic.

References

MF. Rogers and A. Ben-Hur. 2010. The use of gene ontology evidence codes in preventing classifier assessment bias., *Bioinformatics*, 25(9):1173-1177.

M. Krallinger and F. Leitner and C. Rodriguez-Penagos and A. Valencia 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II., *Genome Biol.*, Suppl 2:S1.

S. Orchard and S. Kerrien 2010. Molecular interactions and data standardisation., *Methods Mol Biol.*, 604:309-318