A Lexical Framework for Semantic Annotation of Positive and Negative Regulation Relations in Biomedical Pathways

Sine Zambach*[1], Tine Lassen*[1,2]

[1]Roskilde University, Computer Science,4000 Roskilde, Denmark
[2]Copenhagen Business School, ISV, Dalgas Have 15, 2000 Frederiksberg, Denmark

Email: Sine Zambach*- sz@ruc.dk; Tine Lassen*- tla.isv@cbs.dk;

*Corresponding author

Abstract

Knowledge of regulation relations is widely applied by biomedical researchers in for example experiment design on regulatory pathways and in systems biology. In the work presented here, we analyze in total 28 verbs - and in dept 6 frequently used verbs denoting the regulation relations *regulates*, *positively regulates* and *negatively regulates* through corpus analysis. We propose a formal representation of the acquired knowledge as domain specific semantic frames and the semantic types of the relata of the resulting relationships. We suggest that the acquired knowledge patterns can be used to identify and reason over knowledge represented in texts from the biomedical domain.

1 Introduction

Relations representing positive and negative regulations are widely used in the biomedical domain in systems biology for representation of pathway relations, e.g. [1]. In biomedical texts, verbs denoting regulation relations are used quite frequently, and for information retrieval tools, retrieval of gene-gene regulations has been investigated more or less detailed, cf e.g. [2, 3].

For a more precise retrieval and representation of regulation relations, however, a deeper analysis of the semantics of the different verbs denoting regulation relations can be useful. The knowledge that is acquired from such an analysis can be translated into textual knowledge patterns, or semantic frames, similar to those in the lexical resources FrameNet and VerbNet.

BioFrameNet [4, 5] is a domain-specific extension to FrameNet [6], which is currently being developed. BioFrameNet is concerned with 'intracellular protein transport', and is augmented with domain-specific semantic relations and links to biomedical ontologies. It uses Frame semantics [7] to annotate the meaning of natural language texts, where the frames are expressed in the Description Logic variant of OWL which facilitates inference on knowledge found in texts.

In another related work focusing on regulation, [8,9], a total of 314 abstracts are manually inspected for regulates-relations and ranked patterns of the form e.g. [`Agent`] V-active [`Patient` Action-NN] produced. In addition, "trigger" words concerning regulation from categories of the GRO-ontology [10] are manually identified.

In [11], an approach to indexing biomedical texts by their conceptual content using ontologies, lexico-

syntactic information and semantic role assignment provided by lexical resources is presented. In this approach, the conceptual content of texts is transformed into conceptual feature structures where synonymous but linguistically distinct expressions are given identical representations. This allows for a content-based search which can be useful for document retrieval.

Our aim is to develop a formal semantics of the regulates relation developed from [12], based on a corpus analysis of selected verbs denoting types of 'regulation' within a comparable frame of textual knowledge patterns similar to [11] and [4] as well as an ontological analysis.

Additionally, the focus on agent-patient roles in regulation can support reasoning over additional extracted events as suggested by [13].

2 Corpus Analysis

In this section, we describe the corpus analysis that is a means of identifying the lexico-syntactic patterns that exist for regulatory verbs and their arguments in biomedical texts. These analyses are the basis for a bio-extension of the semantic frames as presented in section 3.

In order to categorize the different types of regulation relations other than through the isa-relation hierarchy in which positive and negative regulations are specializations of regulation, cf. figure 1, we analyze a corpus compiled of biomedical texts or, more specifically, a collection of PubMed abstracts. Through this analysis, we have identified four general types of regulations patterns as outlined below.

Verbs denoting *regulation*, *negative regulation* and *positive regulation* have a somewhat different usage in biomedical texts than they do in general language texts. In order to identify this usage, we created a concordance of all occurrences for a selection of regulatory verbs in a corpus consisting of 40.000 PubMed abstracts. Our search covered the specific verb forms "regulates" (323 occurences) (denoting *regulation*), "inhibits" (781 occurences) (denoting *negative regulation*), "reduces" (699 occurences) (denoting *negative regulation*), "decreases" (1119 occurences) (denoting *negative regulation*), "increases" (3171 occurences) (denoting *positive regulation*), and "stimulates" (372 occurences)(denoting *positive regulation*).

By examining the concordances for these verb forms, we can classify the usage of the examined verbs with respect to types of arguments into four general frame types or patterns (analyzed further in section 3.1). In the patterns presented below, arguments may be of the type *processes* or *substances*. *Substances* can for example be gene products (e.g. proteins and functional RNA) or small molecules and *processes* can for example be glucagon release or glucose transport.

A majority of the identified frame parts are overlapping with the ones identified in [8], however two additional frame parts were identified through our analysis (italicized). The notation form for the frame parts presented below is equal to the one used in [8].

Substances regulate processes. This pattern covers roughly 80% of the occurences of the examined verbs. Example: *"...glp-1 inhibits glucagon release..."* . This correlates with the frame parts in [8] having the forms:

> [`Agent`] V-active [`Patient` Action-NN]
>
> [`Patient` Action-NN] V-passive [`Agent`]
>
> [`Agent`] *is required/essential/involved in* [`Patient` Action-NN]
>
> [Action-NN 'of' `Patient`] *by* [`Agent`]
>
> [`Patient` Action-NN] V-active [`Agent`]
>
> [`Patient` Action-NN] V-active *caused by* [`Agent`]

Substances regulate substances. This pattern covers roughly 10% of the occurences of the examined verbs. For this pattern, the regulated substances are most frequently enzymes. Example: *"...lithium inhibits the enzyme glycogen synthase kinase-3...",* . In terms of [8], the frame parts would be:

> [`Agent`]-Action-JJ [`Patient`]
>
> *[`Agent`] V-active [`Patient`]* (added)

Processes regulate processes. This pattern also covers roughly 10% of the occurences of the examined verbs. Example: *"...nitric oxide pathway regulates pulmonary vascular tone..."* . In terms of [8], the frame parts for this pattern would be:

> [`Agent` Action-NN] V-active *cause* [`Patient` Action-NN]
>
> *[`Agent` Action-NN] V-active [`Patient` Action-NN]* (added)

Processes regulate substances. This pattern covers a minor part of the occurences of the examined verbs. Very few examples of this pattern were found, only in the analysis of the verb *regulates*. Example: *"...Proximal tubular dopamine production regulates basolateral Na-K-ATPase ..."*.

Thus, not many textual instances of regulations have a process on their left hand side, i.e. present the patterns *processes regulate substances* or *processes regulate processes*, but the vast majority of our examples present a pattern where a substance regulates a substance/process. Normally, when regulation relations are represented in biochemical interaction webs such as KEGG [1], they are marked from substance to substance, e.g. "PP1 stimulates GYS" (two gene products). This wording, however, does not reflect the fact that most often the statement is really: "PP1 stimulates *the production* of GYS".

The over-representation of the pattern *substances regulate processes* is also reflected in the number of text patterns found. For example, in [8], 13 patterns are found of which at least 7 are of this form. A semantic discussion of this is given in section 3

In an extended corpus analysis as well as frame-comparison from the resources of FrameNet, VerbNet and WordNet, we notice that some of the verbs representing regulatory relations exhibit a deviant behaviour compared to the identified frame parts. For example, the verbs "increase" and "decrease" often appear in a passive or nominal form, and in these cases, they do not have an expressed agent. We must thus add frame parts such as *[`Patient` Action-NN] V-passive* and *NN 'in' [`Patient` Action-NN]*. This type of linguistic knowledge is important for the outcome of the semantic annotation and eventually for a reasoning over the extracted knowledge.

3 Semantics of Regulation Relations

The results of the corpus analysis as presented in section 2, can be viewed as an extensional definition of regulation relations. However, to be able to perform a reliable semantic annotation of text, there is a need for an understanding of the intensional side of the relations.

3.1 Ontological Types of the Relata

Here, we discuss the connection between the patterns presented in section 2 and the intensional descriptions of the relations as it is described in [12].

Though the proposed transformations are purely formal, they can be useful for a reasoning process as well as for a foundation for a semantic annotation.

In line with [14], we distinguish between ontological types from a top-level ontology, however, we use types from the domain specific top-level ontology of UMLS, the Semantic Network [15]. By using the Semantic Network as our top-ontology, it is possible to identify the ontological types of terms present in the text.

This means that the abovementioned knowledge patterns can be processed such that the semantic constraint, *Substances inhibit processes*, can be incorporated into the knowledge patterns using concepts from the Semantic Network. For example, "Substance(T167)" is a type, having subtypes such as "Amino Acid Peptide or Protein" and "Chemical". Additionally, "Phenomenon or Process(T067)" is type representing "process" having subevents like "Physiologic Function" and "Cell-function".

Since all concepts in the individual UMLS-resources have a direct link into the Semantic Network, this method will make it possible to capture the ontological types of a large number of domain specific terms. The understanding of the main types are as follows:

Substances are, like for instance "continuants" in BFO [16], entities that continue to exists over time and which may undergo changes, contrary to "processes" which are subtypes of "events". Substances are entities that can change and such changes are processes. An example of a substance in our domain, could be an amount of *insulin*, whereas *glycogenesis* is a process.

Substances can regulate other substances or processes, but processes can also regulate other processes or substances. Focusing on for example the relation *regulates* there are therefore four possible relations among individuals combining the two types of relata *substance* and *process*, corresponding to the four general patterns given in section 2. We name these relations $\mathbf{regulates_{ss}}$, $\mathbf{regulates_{sp}}$, $\mathbf{regulates_{ps}}$, and $\mathbf{regulates_{pp}}$, where the subscript "*ss*" means that it is a relation that can only exist between two substances, "*sp*" means that it is a relation that can only exist between a substance and a process,"*ps*" means that it is a relation that can

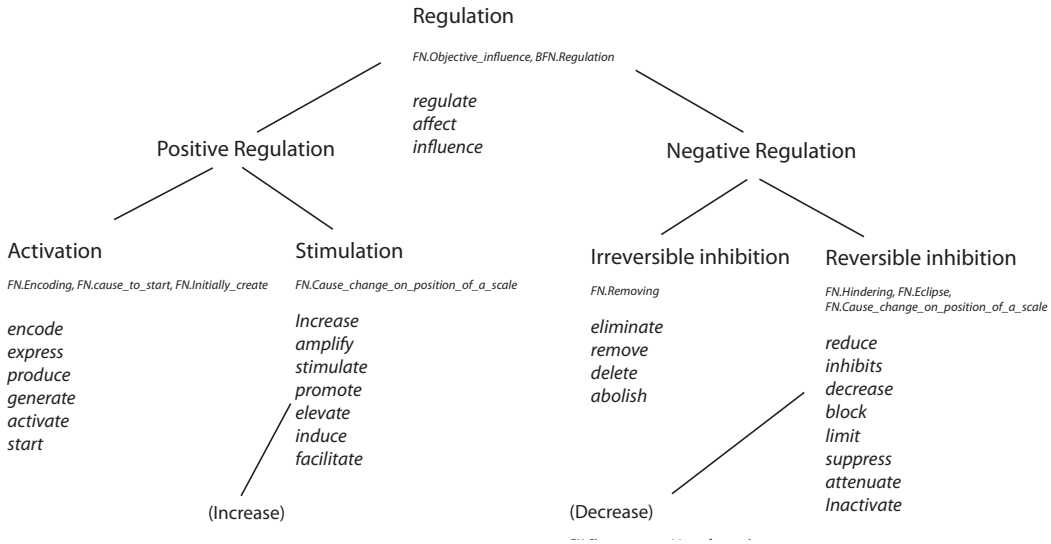

Figure 1: A frame ontology. Based on FrameNet, WordNet and our own corpus analysis, verbs are categorized into a class of this ontology. Other ontologies/terminologies concerning regulation as concepts have been presented in e.g. [10]. Note that the frame BFN.Regulation is our own suggestion.

only exist between a process and a substance, and finally, "*pp*" means that it is a relation that can only exist between two processes given in section 2.

We can formalize the four general types of patterns discussed in section 2. We use $s, s_1, s_2, \ldots$ to range over substances, and $p, p_1, p_2, \ldots$ to range over processes.

Substances regulate substances
$\Rightarrow s_1\ regulates_{ss}\ s_2$
Substances regulate processes
$\Rightarrow s\ regulates_{sp}\ p$
Processes regulate substances
$\Rightarrow p\ regulates_{ps}\ s$
Processes regulate processes
$\Rightarrow p_1\ regulates_{pp}\ p_2$

However, introducing a "production_of" and a "output_of" operator as proposed in [12], makes it possible to reduce these four relations to one, namely $\mathbf{regulates_{sp}}$, as shown in the transformations below.

The "production_of" operator works on a substance s by transforming it to the process that produces s. Similarly the "output_of" operator transforms a process p to the substance that is the output of p. With these operators, the instance relations $\mathbf{regulates_{ss}}$, $\mathbf{regulates_{ps}}$, $\mathbf{regulates_{pp}}$ and $\mathbf{regulates_{sp}}$ can be transformed into one, namely the $\mathbf{regulates_{sp}}$ relation. These transformations are given below

$s_1\ \mathbf{regulates_{ss}}\ s_2$
$\Rightarrow\ s_1\ \mathbf{regulates_{sp}}\ production\_of(s_2)$

$s\ \mathbf{regulates_{sp}}\ p$
$\Rightarrow\ s\ \mathbf{regulates_{sp}}\ p$

$p\ \mathbf{regulates_{ps}}\ s$
$\Rightarrow\ output\_of(p)\ \mathbf{regulates_{sp}}\ production\_of(s)$

$p_1\ \mathbf{regulates_{pp}}\ p_2$
$\Rightarrow\ output\_of(p_1)\ \mathbf{regulates_{sp}}\ p_2$

Additionally, we note that the pattern $s_1\ \mathbf{regulates_{sp}}\ function\_of(s_2)$ denoting a slightly different meaning, namely a regulation by a substance of the function of an enzyme (or another substance), frequently occurs. The specific lexical patterns representing this relation are for example the expressions:

$[\texttt{Agent}]V - active[\texttt{Patient}\ activity/function]$

$[\texttt{Agent}]V - active[activity/function\ of\ \texttt{Patient}]$

Similarily, the *production_of*-operator in

s_1 **regulates$_{\mathbf{sp}}$** *production_of*(s_2), has the expressions:

[Agent]V − active[Patient*production/secretion/transcription/expression/synthesis/release*]

[Agent]V − active[*(the) synthesis/production/secretion/expression/transcription/release of* Patient]

These identified specific lexical patterns expressing the patient, can be of use for extraction of knowledge regarding the pattern for later reasoning as presented briefly earlier in this section.

Reasoning over Knowledge Patterns

These transformations reflect the underlying semantics of verbs denoting regulates relations in biomedical texts. For example, the verb *stimulate* has a usage where it denotes the relation *positively_regulates*[1]: "Insulin **stimulates$_{\mathbf{ss}}$** glycogen", "insulin **stimulates$_{\mathbf{sp}}$** the glycogenesis", and "insulin **stimulates$_{\mathbf{sp}}$** the production of glycogen (through the glygoneonesis)" where the process glycogenesis is equal to "the production of glycogen". Likewise, we can construct the sentences: "beta cell secretion **stimulates$_{\mathbf{pp}}$** glycogenes" that can be transformed to "outout of beta cell secretion **stimulates$_{\mathbf{sp}}$** production of glycogen", where the output of beta cell secretion is insulin.

It may be argued that the relations **regulates$_{\mathbf{ps}}$** and **regulates$_{\mathbf{pp}}$** are not genuine relations since we can question whether it is at all possible for processes to stimulate other processes or substances directly or whether it is rather through their outputs they stimulates.

However, as part of our aim is to be able to identify and annotate instances of regulations in texts, it is important to include the patterns that cover the forms as they actually occur in biomedical texts, and not only as we know them to function. In this cross field between form and meaning, it may be possible both to grab meaningful contents from texts through the semantic roles (e.g. agent and patient) of the relata.

4 Conclusions

In this work we have performed a preliminary investigation of a subset of English verbs denoting regulation relations, and given a more formal account of these relations. We have investigated concordances for 6 verbs; 1 verb denoting *regulation*, 2 verbs denoting *positive regulation* and 3 verbs denoting *negative regulation*, and identified additional textual knowledge patterns compared to former work and the lexical resources FrameNet, WordNet and VerbNet.

In order to achieve a deeper knowledge about the semantics of the studied relations, we initially mapped the relata of the relations into the UMLS Semantic Network. Second we grouped the different verbs denoting regulates relations and their corresponding text-patterns into four different types corresponding to the types of their relata.

Finally, we gave a formal description of the four observed types of regulation relations and transformed these into one. Using the operators "production_of", "output_of", and "function_of", the four observed types of regulation relations, **regulates$_{\mathbf{ss}}$**, **regulates$_{\mathbf{sp}}$**, **regulates$_{\mathbf{ps}}$**, and **regulates$_{\mathbf{pp}}$**, can be transformed into one, namely the **regulates$_{\mathbf{sp}}$** relation.

We stress the importance of identifying textual forms of the relations as knowlege patterns as they actually occur in biomedical texts, even if we know that a given pattern does not reflect how regulations function in reality as we know them to function.

The semantic patterns that are identified through our corpus analysis, can form a background for further knowledge extraction, where for example text is automatically annotated by use of the patterns and subsequently fed to a machine learning algorithm for identification of new patterns (in line with [9]). This could be a part of an automatic semantic annotation for semantic based information retrieval. The semantic roles that are gained by such an attempt, should be precise enough for being part of a knowledge base that can be enriched with reasoning rules.

Additionally, through a deeper linguistic analysis, these parts can be part of the basis for a domain specific FrameNet describing regulatory events (an extended BioFrameNet in line with the vision of [5]).

[1]This example is also used in [12].

References

1. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res.* 2000, **28**:27–30.
2. Chen H, Sharp BM: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5**:147.
3. Hoffmann R: **Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds**. *Curr Protoc Bioinformatics* 2007, **Chapter 1**:Unit1.16.
4. Dolbey A, Ellsworth M, Scheffczyk J: **Bioframenet: A domain-specific framenet extension with links to biomedical ontologies**. *Proceedings of KR-MED* 2006, :87–94.
5. Dolbey AE: **BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology**. *PhD thesis*, University of California, Berkeley 2009.
6. Fillmore C, Johnson C, Petruck M: **Background to Framenet**. *International Journal of Lexicography* September 2003, **16**:235–250(16).
7. Fillmore CJ: **Frame semantics and the nature of language**. In *Origins and evolution of language and speech*. Edited by Harnad S, Academy of Sciences 1976:155–202.
8. Buyko E, Beisswanger E, Hahn U: **Testing Different ACE-Style Feature Sets for the Extraction of Gene Regulation Relations from MEDLINE Abstracts**. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*. Edited by Salakoski T, Rebholz-Schuhmann D, Pyysalo S, Turku Centre for Computer Science (TUCS) 2008:21–28.
9. Hahn U, Tomanek K, Buyko E, Kim Jj, Rebholz-Schuhmann D: **How feasible and robust is the automatic extraction of gene regulation events?: a cross-method evaluation under lab and real-life conditions**. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, Morristown, NJ, USA: Association for Computational Linguistics 2009:37–45.
10. Beisswanger E, Lee V, Kim JJ, Rebholz-Schuhmann D, Splendiani A, Dameron O, Schulz S, Hahn U: **Gene Regulation Ontology (GRO): design principles and use cases.** *Stud Health Technol Inform* 2008, **136**(NIL):9–14.
11. T Andreasen PAJ H Bulskov, Lassen T: **Conceptual Indexing of Text Using Ontologies and Lexical Resources**. In *Proceedings of the Eigth International Conference on Flexible Query Answering Systems*, Springer 2009.
12. Zambach S, Hansen J: **Logical knowledge representation of regulatory relations in biomedical pathways**. In *Proceedings of ITBAM-DEXA*, Lecture Notes in Computer Science, Springer, LNCS 2010.
13. Zambach S: **A formal framework on the semantics of regulatory relations and their presence as verbs in biomedical texts**. In *Proceedings of the Eigth International Conference on Flexible Query Answering Systems*, Lecture Notes in Computer Science, Springer 2009:443–452.
14. Smith B, Ceusters W, Klagges B, Ko hler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: **Relations in biomedical ontologies**. *Genome Biol.* 2005, **6**:R46.
15. Mccray AT: **An Upper-Level Ontology for the Biomedical Domain**. *Comp Funct Genom* 2003, **9**(4):80–4, [http://citeseer.ist.psu.edu/706880.html;http://lhncbc.nlm.nih.gov:80/lhc/docs/published/2003/pub2003023.pdf].
16. **Basic Formal Ontology.**[http://ontology.buffalo.edu/bfo/].