Regulatory Relations Represented in Logics and Biomedical Texts

by
Sine Zambach

Supervised by Troels Andreasen

A dissertation submitted in partial satisfaction of the requirements for the PhD degree

Department of Communication, Business and Information Technologies,
Roskilde University

February 3, 2012

Contents

List of Figures

List of Tables

English Abstract

Regulatory networks are used for simple modeling of varying complexity, for example within biology, economics and other fields that apply dynamic systems.

In biomedicine, regulatory networks are widely used to model regulatory pathways, which, in short, are characterized by processes containing gene products and smaller molecules that regulate each other through different mechanisms through different paths. The relations between the building blocks of these networks are typically modeled either very expressively or very simply in graphs in information systems.

The focus of this dissertation is the biomedical semantics of *regulates* relations, i.e. *positively regulates*, *negatively regulates* and *regulates*, of which is assumed to be a super relation of the first two.

This thesis discusses an initial framework for knowledge representation based on logics, carries out a corpus analysis on the verbs representing *regulation (positively* and *negatively)* and defines four rules of reasoning. Compositions within class relationship logic have also been explored.

One of the main goals of this dissertation is to form a foundational basis of knowledge on regulation that contributes to the further development of information services on regulatory events within biomedicine.

Dansk Resumé

Regulatoriske netværk bruges til computermodeller af varierende kompleksitet, for eksempel inden for biologi, økonomi og andre fagområder, der anvender dynamiske systemer.

I biomedicin er regulatoriske netværk i vid udstrækning anvendt til at modellere regulatoriske pathways. Disse er groft sagt karakteriseret ved substanser (genprodukter og mindre molekyler) som regulerer hinanden gennem forskellige mekanismer og processer via forskellige veje. Relationerne, eller vejene, i disse netværk er typisk modelleret enten meget matematisk sofistikeret eller som simple grafer.

I dette arbejde er fokus på biomedicinsk semantik for relationerne *positiv regulering* og *negativ regulering* såvel som *neutral regulering*, som vi antager er et overbegreb til de to andre. Vi kalder de tre relationer "regulatoriske relationer".

Der diskuteres et grundlag for vidensrepræsentation som er baseret på logik, og en korpusanalyse vedrørende både hyppigheden af verber der repræsenterer *regulering* (*positiv* og *negativ*). Desuden er fire ræssoneringsregler formaliseret baseret på kompositioner af relationer mellem klasser.

Denne afhandling skal danne et grundlag for viden om regulering med det formål at udvikle it-services baseret på regulatoriske events inden for biomedicin.

Chapter 1

Introduction

In the relatively new field of biomedical informatics, which is not represented by a vast literature, the introductions in "calls for papers" for conferences and journals have proven to be useful to my research. Although I do not quote the introductions, they provide inspiration and keep one abreast of what issues are currently of interest in the field.

A call for papers in a new journal, *Journal of Biomedical Semantics*, focuses on the importance of a biomedical practice that investigates and develops devices for interdisciplinary cooperation in the fields of informatics and biomedicine:[1]

> *"Research in biology and biomedicine relies on various types of biomedical data, information and knowledge, represented in databases with experimental and/or curated data, ontologies, literature, taxonomies, etc. Semantics is essential for accessing, integrating and analyzing such data. The ability to explicitly extract, assign and manage semantic representations is crucial for making computational approaches in the biomedical domain productive for a large user community."*

This focus is relevant for my research and highlights my motivation for developing the above ideas further.

The aim of this thesis is to investigate regulation among molecules and the opportunities available for developing logic-based information systems that can provide (new) knowledge about regulatory networks in a biomedical context. This development is based on data and information transformed into knowledge to achieve the ability to "explicitly extract, assign and manage semantic representations" as *Journal of Biomedical Semantics* states.

Although there are various shades of difference between information and knowledge, as will be explained in section 1.1, one common goal shared by researchers in biomedical informatics is making data and information available

[1] http://www.jbiomedsem.com/info/about/, August 2011

for biomedical researchers and physicians. As stated in [116], biomedical ontology development must work to provide practitioners with the knowledge they need to meaningfully utilize ontologies.

In addressing some of these issues, foundations within ontologies and applications are treated to meet the need for reorganizing (in the sense of developing bioinformatics software in a new way) and to investigate the heterogeneous area of biomedical informatics.

1.1 Research questions

The overall research questions are:

> *What are the semantics of "regulates" as a relation in molecular biology?*
>
> - *Can a formal description based on first-order logics create the basis for reasoning over regulates relations?*
> - *How are regulatory events represented as relationship triples in biomedical texts?*
> - *How can relationship triples and the formalization of regulates be tractably utilized in hypothesis testing?*

In brief, how can formal knowledge of regulation within molecular biology provide useful reasoning for hypothesis testing in a practical and constructive manner?

The objective is to collect and investigate both the *intensional* meaning of regulates as a relation as understood in biochemistry and the *extensions* of the (conceptualized) relation as included here in multiple text examples.

An additional focus is on the semantics of regulation and how they can be utilized for reasoning rather than solely using "knowledge," which can be interpreted in many ways, as will be briefly discussed later.

Another topic is the balance between tractability and expressiveness in formal implementation, which is a general problem in most sub fields of knowledge representation.

The subsequent sections introduce the central research concepts of this thesis.

1.1.1 Bioinformatics and bio-knowledge

Mentioned for the first time in 1984 [18], the field of bioinformatics began growing slowly in the late 1980s. The discipline continued throughout the 1990s, the term being widely used, for example, in the NCBI database [131], which was established in 1997, slightly more than ten years after the human genome project was initiated in 1986. Since the late 1990s, bioinformatics research has undergone even further development and the amount of sequence data being published has been expanding exponentially.

A niche of bioinformatics, the study of biological and medical knowledge, often referred to as biomedical informatics, biomedical ontology or bio-knowledge, has also expanded in the past decade.

Whereas bioinformatics mainly refers to the analysis of gene and protein sequences, but also systems biology, biomedical informatics is often used as a joint concept for the foundation of computational disciplines such as ontology research, artificial intelligence and natural language processing as

it operates in the biomedical domain. A more appropriate name could be biomedical knowledge engineering, as the aforementioned fields are really aspects of knowledge engineering. Since biomedical informatics is commonly used by the general public, this term is also employed in this thesis.

The different disciplines within biomedical informatics deal with knowledge or semantics and emerged on a large scale in the medical expert systems for physicians in the 1980s. In 1990, the ontology repository UMLS was launched [82]. In 1997, SNOMED RT [72], a reference terminology for health care, was introduced, and in 2000, the Gene Ontology was introduced [14], though ontology was not a focus in its early stages. In the past five years, biomedical informatics societies have expanded with organizations such as the Open Biological and Biomedical Ontologies (OBO) Foundry [118], the National Center for Biomedical Ontology [115] and the International Conference on Biomedical Ontologies (ICBO) conferences.[2]

Launched in 2001, the *Journal of Biomedical Informatics* has illustrated the differences between biomedical informatics methods and other disciplines such as bioinformatics and clinical informatics, see figure 1.1.[3] Biomedical informatics creates the foundation, background knowledge and principles (or knowledge engineering) for applied disciplines such as bioinformatics and clinical informatics. Thus, knowledge engineering plays an important role in biomedical informatics and the next section will introduce how the concepts of knowledge, information, and data are used in this thesis.

1.1.2 Knowledge, Information and Data

After introducing the concepts of biomedical informatics, the ways in which data, information and knowledge are distinct from one another will be discussed. These concepts are central to this dissertation, where one of the objectives is to provide knowledge from data and information that can be used and processed by the end user, typically a biomedical researcher.

Data can be understood as (structured) raw pieces of material or symbols for facts that exist but have no meaning individually. Examples of data, such as a graph of points, are incomprehensible without a legend or explanation. For instance, the introduction of a nucleic acid sequence (e.g. *atgctgctttggaa*) without annotation or definition would be meaningless [47].

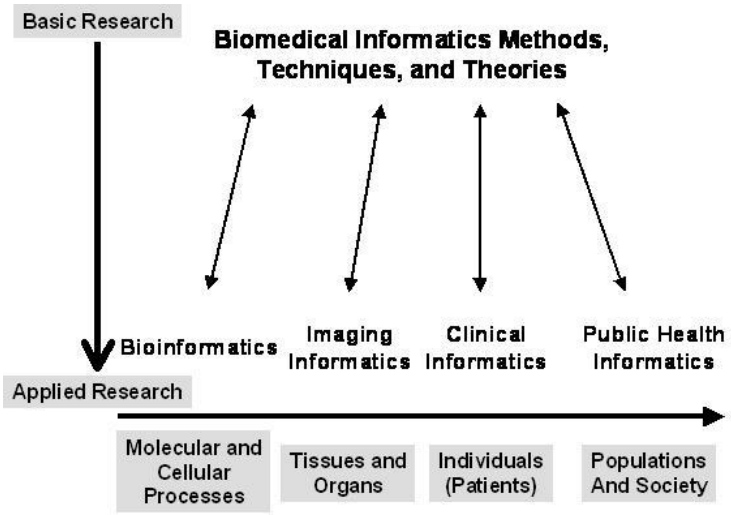
Information translates data providing context and meaning. A graph with an explanation is information. A relational database scheme provides information about the data in a database, and a gene sequence with information on which structures could be genes is information. A simple, popular way (General Definition of Information, GDI [47]) to define information is:

$$information = data + semantics. \tag{1.1}$$

[2] www.icbo.org

[3] From the website of Journal of biomedical informatics, ees.elsevier.com/jbi/

Figure 1.1: Illustration of biomedical informatics as understood by the *Journal of Biomedical Informatics*.



In this equation, semantics refers to the underlying meanings of a pattern. Note that semantics can be understood both concretely as the meaning of a word, but also as a more abstract or formal notion, where the semantics of algebra or language facilitate the logical manipulation of (ideal) sentences [128]. In e.g. Fillmore's frame semantics [46], a basis exists for combining the semantics of natural language with a logical formalization, i.e. for being able to manipulate and characterize *frames* (collections of semantically similar lexical items with similar contexts) logically.

Knowledge, as used in this thesis, requires a subject or a group of subjects. Knowledge is information that can be retrieved and is generally believed to be true by a subject or a group of subjects [47, 120]. Moreover, knowledge offers the opportunity to have explanations and entailments, since, if believed to be true, entailments based on knowledge are the desired goal.

Information in PubMed is turned into knowledge by the fact that it is a verified database of peer-reviewed articles that society generally believes are true, and by the semantic patterns that will be presented later. From an information scientific perspective, an ontology is a representation of knowledge (Introduction, [120]).

According to the linguistic and logics approach of this thesis, a *syntactical text-pattern* cannot be a *lexico-semantic pattern* [69] before it contains what is required for: a) linking terms to the semantics/concepts in the ontology; and b) reasoning over the concepts (formalization).

When semantic patterns are used in systems of axioms and assertions

that are either specific or general for the domain,[4] and they are reusable, they equal a *knowledge pattern* as suggested in [34]. Table 5.2 sums up the terms concerning different patterns in texts.

In *knowledge engineering*, an intensional definition of a concept plays a more prominent role than an extensional one, since it can help identify and classify objects with respect to their properties.

On the other hand, extensional definitions create a base for collecting a catalog or database of individuals corresponding to a concept (e.g. section 2.6 in [123]). The extensional understanding of a concept may be important in order to properly utilize *information* and *data* corresponding to the concept. The bottom-up nature of an extensional definition can help developers capture information and map it into the concepts of interest within specific domains.

In the technology of so-called *surface text patterns* [113], multiple paraphrases concerning the same type of events are used for efficient machine learning. This is a bottom-up procedure and it is extensional in the sense that the aim is to find the expressions of a problem and then to provide a solution (in the case of [113], query and answering). Although the patterns alone cannot help to develop axioms and assertions for reasoning, they are necessary to extract data and information in, for example, databases and article abstracts, and for initial analysis.

On the other hand, a rather top-down procedure is the use of semantic text patterns based on ontological types, or what is also called lexico-semantic patterns, and can be powerful in knowledge extraction in combination with surface text patterns. These patterns are more formal and combine the background knowledge as domain constraints and linguistic constructs [69] similar to the approaches described in chapters 4 and 5 and they suggests multiple paraphrases of semantic text patterns. Table 5.2 providing an overview of the use of annotation concepts such as lexico-semamtic patterns, frame parts, etc.

Thus, to me, intensional and extensional ontology/semantics are two sides of the same coin. This thesis is an attempt to formalize knowledge of regulation on the basis of information and in the understanding that while information is something you can retrieve, knowledge is additionally something you can reason over within a domain (in a reliable way).

1.1.3 Knowledge representation

All data, information and knowledge have a representation that can be informal in spoken and written natural language; implemented in a database

[4]Notice that this interpretation of knowledge does not need to be deterministic in a Boolean "True or False" sense. It can as well be probabilistic or fuzzy reflecting deviations and uncertainties within the real world.

or semantically tagged text or represented using logical formalisms. Any of such representations are often called knowledge representations (KR).

The fundamental goal of KR is typically to represent knowledge to facilitate inferences and reasoning from knowledge [123]. On a less abstract level, Woods [130] suggested in 1975 that KR languages or formalisms should "(...)unambiguously represent any interpretation of a sentence (logical adequacy), have a method for translating from natural language to that representation, and must be usable for reasoning" [130].

However, a KR is not in itself a data structure but the underlying semantics. Due to [39], KR has at least five roles:

1. As a surrogate for truth (this is a premise for knowledge representation rather than a defining role).

2. As a set of ontological commitments (with the constraint that the view or norm, added to concepts has an influence on the KR).

3. A fragmentary theory of intelligent reasoning (a language or formalism from which new facts can be inferred).

4. A medium for efficient computation (i.e. a tractable formalism).

5. A medium for human expression (i.e. a formalism that has an adequately high level of expressiveness).

In summary, KRs have many roles. The most basic being that it should be possible for a computer to interpret them and that they should provide useful knowledge for the end user.

As applied here, KRs are mainly used to gain an opportunity to reason over knowledge of interest within a given domain, and with appropriate expressiveness. This opportunity for reasoning within traditional formalisms, like first-order logics and fragments of this, as well as within the discipline of extracting knowledge (or information) from texts is explored.

1.1.4 Motivation and process

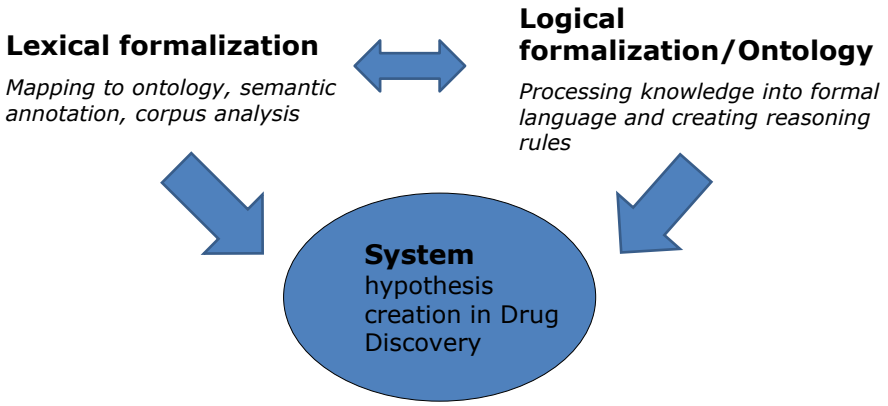
Based on this understanding of knowledge and KR, and inspired by a knowledge engineering approach, the work flow of this thesis can be described as follows:

- First, a basic understanding of KR and ontologies is necessary.

- As a result, semantics and formalisms of knowledge must be considered.

- The domain of discourse must be analyzed to obtain background knowledge for the KRs.

Figure 1.2: Knowledge representation - from semantics to system.



- Practical modeling of classes and relations needs to be done using real information; taking into consideration the reasoning and inferences from the representation's output is essential.
- Finally, implementation perspectives should be proposed.

The process is not necessarily as linear as it is described here, and has not been linear in the work of this dissertation. Considerations concerning reasoning and an analysis of formal semantics are intrinsically connected, which is illustrated in the figure 1.2.

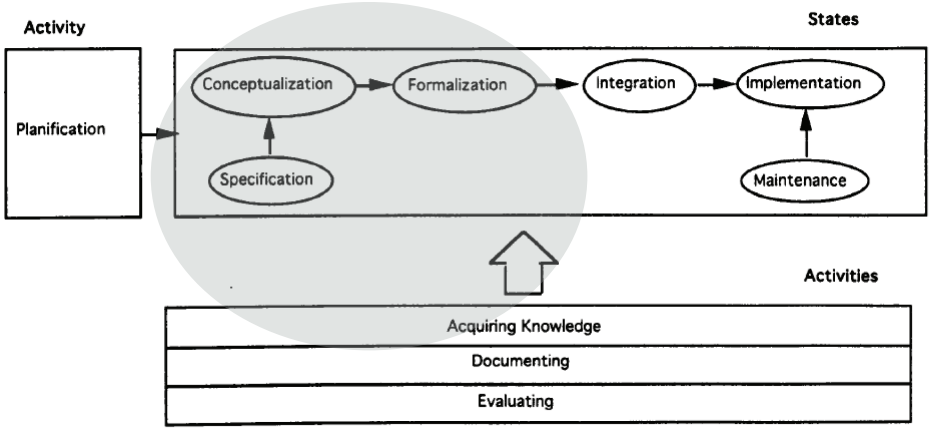
In addition, having an idea about the system's outcome is central. Implementation options might be as important as finding the most correct semantics. Implementation options are also interlinked with the process of determining what to consider concerning formalism.

The work done on the reasoning process of the domain (section 4.5 [134, 135]) in this dissertation actually took place prior to the work carried out on the deeper domain (section 4) and the formal semantics (section 4.4).

This process described above is parallel to a process proposed in METHONTOLOGY, an ontology engineering method [52, 44].

The task of the ontology should be figured out, for example, before building and formalizing it. Figure 1.3 illustrates a modeling focus based on METHONTOLOGY.

Figure 1.3: Modeling focus based on METHONTOLOGY [44]. The shaded area covers the main contributions of this thesis within the framework of ontology development.



1.2 Related works

Knowledge representation and reasoning in the field of biomedicine continues to expand. This section introduces biomedical ontologies for regulation, an ontology of formal (biomedical) relations and, finally, various information retrieval tools.

1.2.1 Regulation ontologies

The majority of the remainder of this dissertation examines ontology of regulation and this section presents works that have already been accomplished in this area.

A movement towards the formalization of biomedical ontologies and the relations the ontologies contain has progressed over the last decade. For example: the widely used Gene Ontology [51, 14] has been ontologically "cleaned up"; initiatives like OBO provide a framework for formal relations and cooperative ontology modeling [118, 23]; and Role Ontology deals with properties for relations [119]. In pathway modeling, especially for regulation, ontologies and formalized systems like the Gene Regulation Ontology [21], Ecocyc/Metacyc [30] and Pathway Logics [43] have recommended a variety of ways to approach logic representation.

In Gene Regulation Ontology (GRO), the domain of regulation is modeled within the ontology. The purpose is to formalize concepts for gene

Table 1.1: Concepts with the so called *negative* and *positive polarity* in GRO using description logic notation [21].

Concept	$\mathcal{DL}$**-definition**	**Natural Language Definition**
Increase	$\sqsubseteq Process$ $\sqsubseteq \exists hasQuality.PositivePolarity$	A process of becoming larger, more numerous, more important, or more likely.
Decrease	$\sqsubseteq Process$ $\sqsubseteq \exists hasQuality.NegativePolarity$	A process of becoming smaller, less numerous, less important, or less likely.
Positive Regulation	$\equiv RegulatoryProcess$ $\equiv \exists hasQuality.PositivePolarity$	Any process that activates or increases the frequency, rate or extent of a biological process, function or phenomenon.
Negative Regulation	$\equiv RegulatoryProcess$ $\equiv \exists hasQuality.NegativePolarity$	Any process that stops, prevents or reduces the frequency, rate or extent of a biological process, function or phenomenon.
Activation	$\sqsubseteq RegulatoryProcess$ $\sqsubseteq \exists hasQuality.PositivePolarity$	Any process that activates, maintains or increases the frequency, rate or extent of an action.
Inhibition	$\sqsubseteq RegulatoryProcess$ $\sqsubseteq \exists hasQuality.NegativePolarity$	Any process that stops, prevents or reduces the frequency, rate or extent of an action.

regulation used in, for example, the Gene Ontology [14].[5] Within the ontology, positive and negative regulation and similar terms are distinguished by means of the delineations: *hasQuality* relation with the values *NegativePolarity* and *PositivePolarity*. GRO distinguish between e.g. *NegativeRegulation* and *Decrease*, which will be examined more closely in section 5.4.

Table 1.1 shows concepts with regulatory semantics including their formalization in description logics($\mathcal{DL}$).[6] These ontologies have proven to be useful in making the move from regulation as a concept to regulates as a conceptual relation.

[5]Gene Ontology uses regulation within some concepts and as a relation [51] though they have not modeled it separately.

[6]The original definition of *NegativeRegulation* was "The process by which a cell decreases the number of a cellular component, such as RNA or protein, in response to an external variable". However, since the rest of the definitions was similar the corresponding opposite polarity, the definition is changed into the one represented in this table. The definitions of *Increase* and *Decrease* were denoted as "comments" in the GRO-owl file. The definition of *PositiveRegulation* and *NegativeRegulation* are similar to the definitions of *Positive regulation of biological process* and *Negative regulation of biological process* in GO [14].

1.2.2 Formal biomedical relations

This section describes some outcomes of the semantic analysis of relations and their uses in biomedical ontologies and semantic webs, primarily based on references: [51, 119, 23]

Role Ontology (RO) is an initiative to collect a number of relations that are general enough to cover the biomedical domain and examples can be found in table 1.2. The relations in this ontology are general and not specific for the biomedical domain and do not contain *regulates* at this point. Nonetheless, they were selected by a collaboration between domain experts and ontology engineers [119].

1.2.3 BioFrames and text patterns

BioFrameNet [41, 42] is a domain-specific extension to FrameNet [71], which is currently being developed. BioFrameNet covers intracellular protein transport and is augmented with domain-specific semantic relations and links to biomedical ontologies (e.g. Gene Ontology). This approach uses frame semantics [46] to annotate the meaning of natural language texts, where the frames are expressed in the $\mathcal{DL}$ variant of Web Ontology Language(OWL), which facilitates inference on knowledge found in texts.

Regulatory events. In another related work, [27] manually inspect 314 abstracts for regulates-relation and make a rank of patterns on the form *[Agent]V-active[Patient Action-NN]* (in other litterature, e.g. [5], the structure is called *agent-regulates-theme,* but the essence is the same). In addition, the authors manually identify "trigger" words concerned with regulation from categories in the Gene Regulation Ontology (GRO) [21]. The notation of the frames is equal to that of [27].

In the work of [27, 60], the paths are used for machine learning and a semantically annotated corpus [28]. Interestingly, there were 9000 negative instances and 1135 positive instances in the training data. In the verb frequencies [134] demonstrated in this dissertation, positive verbs are more common.

Dolby 2009 [42] identified frames for molecular transport events, which were termed BioFrameNet, intending to be integrated as a domain specific FrameNet module. Similarly, this dissertation, as an extension of BioFrameNet, also includes regulatory events.

Other works concerning events and regulatory events focus on relations and their arguments (or *relata* as they are often called in this thesis) as stated in [5]. Many of these techniques focus mainly on protein-protein interactions, including the Gene Regulation Event Corpus(GREC) [127] and GeneReg [28], which are mapped to GRO [21].

Table 1.2: Part of the Role Ontology [119]. Foundational relations are included in the Role Ontology (in the OBO-project [2]).

Name	Properties	Formal definition
is_a	[transitive] [reflexive] [anti-symmetric]	For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P'.
part_of	[transitive] [reflexive] [anti-symmetric]	For continuants: C part_of C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t, and c *part_of* c' at t. For processes: P part_of P' if and only if: given any p that instantiates P at a time t, there is some p' such that p' instantiates P' at time t, and p *part_of* p' at t. (Here *part_of* is the instance-level part-relation.)
integral_part_of	[transitive] [reflexive] [anti-symmetric]	C integral_part_of C' if and only if: C part_of C' AND C' has_part C
proper_part_of	[transitive]	As for part_of, with the additional constraint that subject and object are distinct
located_in	[transitive] [reflexive]	C located_in C' if and only if: given any c that instantiates C at a time t, there is some c' such that: c' instantiates C' at time t and c *located_in* c'. (Here *located_in* is the instance-level location relation.)
contained_in		C contained_in C' if and only if: given any instance c that instantiates C at a time t, there is some c' such that: c' instantiates C' at time t and c located_in c' at t, and it is not the case that c *overlaps* c' at t. (c' is a conduit or cavity.)
adjacent_to		C adjacent to C' if and only if: given any instance c that instantiates C at a time t, there is some c' such that: c' instantiates C' at time t and c and c' are in spatial proximity
has_participant		P has_participant C if and only if: given any process p that instantiates P there is some continuant c, and some time t, such that: c instantiates C at t and c participates in p at t
has_agent		As for has_participant, but with the additional condition that the component instance is causally active in the relevant process
...	...	...

Additionally, the GENIA corpus [76], which contains 97 annotated full-text articles, have been developed, but do not focus on regulatory events, though they contain information on these as well as other issues.

1.2.4 Related IR-tools

The biomedical literature database, PubMed is growing by approximately 50 papers a day and studies have shown that approximately 30 percent of the co-occurred protein pairs in the abstracts interact [5].

A few information retrieval tools that use information on regulative events also exist. Introduced below are iHop [65][7] and Chilibot [32],[8] which have each been available since 2004. They both focus on protein-protein interaction extraction and use semantic annotation.

Chilibot contains information on regulation, where the author categorizes the regulation relations into positive, negative and neutral. Thus, the system does not utilize a hierarchy of relations or a fine-grained differentiation of multiple verbs representing regulative events, but has a rough differentiation and a semantic retrieval of regulatory interactions. Unfortunately, this tool has not been developed or updated since 2007.

iHop is a large-scale information retrieval tool that extracts interaction information among proteins in Medline abstracts. This tool is updated daily and offers searches through texts on specific organisms. The output, however, is a web of interactions without any semantics of the interaction type. Figure 1.4 shows a screen dump of a text with the regulation "trigger" verbs highlighted. There is no attempt so far to parse the sentence with respect to the semantics of the interaction form.

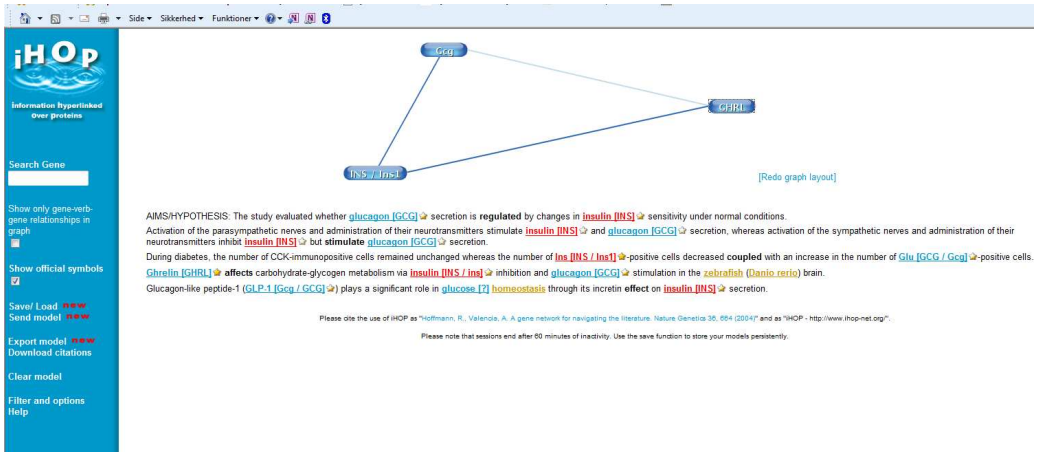
None of the web tools employ reason over interaction for the extracted texts. This could be a case of not introducing too much uncertainty or avoiding a careful semantic curation of the text patterns that correspond to the interactions.

Within entity recognition, the tool Reflect [106] supplies semantic tagging of substances such as proteins and chemicals, together with an access to the Uniprot database. The unique MetaMap tool facilitates the mapping of biomedical text into the UMLS Metathesaurus [82] as well as the UMLS Semantic Network [92]. This is done by parsing the text and tagging it with terms from UMLS.

Finally, web services like Reactome [91][9] and STITCH [78][10] contain simple interaction information in pathways based on several interaction databases.

[7] http://www.ihop-net.org/
[8] http://www.chilibot.net/
[9] http://www.reactome.org/
[10] http://stitch.embl.de/

Figure 1.4: iHOP screen dump [65].



These, however, do not utilize the information that is received from texts and they also will not predict the polarity of interaction (i.e. will this molecule inhibit or stimulate the process).

1.3 Foundations and contributions

The focus and problem area for discussion in this dissertation is mainly within the domain of regulation within biochemistry and regulatory networks, and specifically regulates as relation. As a specific focus within regulatory events the "agent" and "patient" roles of the relata surrounding the relation as represented in texts are investigated. This forms the link between the corpus work, done in section 5, and the formal work presented in section 4.4. In short, the focus has been on how roles/relata are represented in texts and knowledge bases and how this information can be useful to reasoning and retrieval.

The theoretical basis for this dissertation is knowledge representation, with the subcategories molecular biology (domain knowledge), mathematics and philosophy (logical knowledge representation, including formal ontologies, semantic analysis and reasoning), computational linguistics (terminology modeling, corpus analysis and lexical semantics), and information science/computer science (information retrieval, system functionalities and implementations).

Previously published work [4, 9, 38, 134, 136, 138, 139, 135, 133] has concentrated on logical representation, reasoning, modeling and textual representation in domain corpora of regulation in biomedicine. Some of the works are very domain specific as [134, 136, 138, 135, 133], whereas a few publications focus on formal methods, with a domain perspective [4, 9, 38, 139].

The main contribution of this thesis is the logical semantic analysis in [136] in sections 4.3 and 4.4; the formalization of reasoning rules and corpus analysis in [134] in sections 4.5 and 5.3.1; and an extension of [138] in section 5.4, which contains corpus analysis and lexical frames.

Finally, the main contributions of the author for the SIABO project which is presented in section 6.1.1, are the modeling of a micro-ontology (as described in section 5.1) developments in the representation formalism class relationship logic ($\mathcal{CRL}$) (which is described in section 3.1.3 and the contribution to compositions is presented in section 3.2), and the investigation of regulatory events in texts for semantic annotation. The SIABO project is presented in [9] and [10].

1.3.1 Thesis structure

Although structured as a monograph, this thesis includes sections based on published papers and/or revisions of published papers. When this is the case, it is noted. The end of each chapter contains a summary, and the end of the thesis contains a list of contributions.

Chapters 2 and 3 present the dissertation's formal foundations and cover the ontologies, knowledge representations and formalisms used to represent the knowledge. Chapter 4 presents the domain of regulation within molecu-

lar biology and formalizes certain aspects of molecular regulation using the formlisms found in chapter 3.

The analysis carried out in chapter 5 also encompasses a linguistic analysis of a corpus, concerning both terminological modeling (section 5.1) and an analysis of different verbs representing regulates (section 5.3). The aim is to investigate the use of regulation in the biomedical language as well as to support the extraction of predicate arguments and later reasoning over these arguments.

Chapter 6 presents semantic information retrieval utilizing the semantic annotation outcome of, for example, the semantic types in chapter 5. Additionally, one of the experimental prototypes that have been developed simultaneously with the work presented in section 4.4 and section 4.5 is presented.

Contributions

The formalization of the regulates relation in [136] (section 4.4) and the ontology modeling in [139, 38] (section 5) is concentrated on the intensional semantics of regulation and its sub-types.

The first work [136], a semantic analysis of regulates as relation, contributed the logical formalization of regulates as a relation among classes, a suggestion of the granularity of individuals within the class, and what types the relata consist of. The next work [139, 38], modeled negative regulation (inhibition) and its sub-types within enzyme chemistry, and contributed a fine-grained view on inhibition along with a modeling procedure for combining terminology-modeling principles (as [86]) with the popular web ontology language OWL [53]. The author's contribution to the work on the SIABO project presented in [9] was also concerned with domain ontology modeling.

A work on minimum corpora [62] (similar to the approach in [9]), also contributes methodology on information retrieval, as well as domain analysis and research ideas.

An extensional view is taken in [133] and parts of [134, 135] which analyze frequencies in biomedical corpora. Additionally, the work on semantic patterns in [138] is a first step toward capturing the extension of concepts from the texts and mapping them into the intensional relationships identified.

The papers on the reasoning rules of positive and negative inhibition, [134, 135], are a step towards knowledge extraction or knowledge retrieval comprising reasoning rules for regulations. The contribution within knowledge extraction is the reasoning rules presented describing how to utilize a formal implementation of the regulatory relations in hypothesis testing. Additionally, a minor contribution within logic/formalism developmental work considered the kind of $\mathcal{CRL}$ relations that can logically be combined in compositions (complex role inclusions, relational closures, reasoning rules) [4].

Finally, a discussion follows in which the topics present in multiple chapters are assembled and future work suggested.

Additional files and a PDF version of this dissertation are available here: *www.ruc.dk/~sz/Regrel/thesis*

Other works

Besides the article and work already presented, the idea behind the paper "Corporate Social Responsibility in Enterprise Systems"[137], concerning environmental IT within a company's existing enterprise system, has been presented. This research is not directly linked to the main focus and thus is not included in this thesis.

Acknowledgments

This dissertation would not have been possible without the support of various people.

First, I would like to thank the SIABO project and the SIABO group (with special thanks to my supervisor Troels Andreasen, Tine Lassen, Bodil N. Madsen and Jørgen F. Nilsson) for their cooperation. I am also grateful to Jørgen F. Nilsson for the introduction to formal semantics and the applications for this subject.

Next, I would like to thank my colleagues Jens Ulrik Hansen, Mai Ajspur (who also drew the figures in section 3.2), Christian Theil Have, Ture Damhus and Peder Olesen Larsen for beneficial co-authorship. Special thanks to Jens Ulrik Hansen and Ole Torp Lassen for co-organizing e.g. a description logics study group, logic seminars in Saarbrucken and Leuven, ESSLLI 2010 and various social events.

I would also like to express my gratitude to the following for their input and important discussions: Matthieu Petit, Uwe Schwarts, Christoph Benzmüller, Søren Mørch, Claus Desler, Kasper Risager, Zenia Worm Francker and Philip Holst. Thanks are also due Vincent Hendricks for co-organizing ESSLLI 2010 in Copenhagen and the Bioinformatic Center at the University of Copenhagen for providing GEO extracts.

During my exchange I attended the Department of Disease Systems Biology headed by Søren Brunak and Lars Juhl Jensen. Thanks for allowing me in the group and for the feedback on the applied and biomedical aspect of my thesis. Also thanks to the department of Computational Linguistics at Copenhagen Business School, which has hosted me for a couple of months.

Finally, I would like to thank my family, especially my children, who taught me the importance of being incredibly efficient with my time, and my husband, who has been exceptionally supportive throughout.

Chapter 2

Ontologies and Knowledge Representation

Ontologies are some of the main elements in the field of knowledge representation (KR). An ontology should describe the world and, in doing so, can be an important tool for the knowledge engineer who works with representing knowledge (about the world).

There are several methodological issues to consider concerning ontology: the degree of structure, what it should contain, how close it needs to be to the real world and whether it should reflect mental concepts. An important consideration to be made before modeling or utilizing ontological resources is, therefore, determining the task to be solved (e.g. [36]).

On the one hand, if the ontology is for reasoning and knowledge exploration within the ontology, it is not unimportant what methodological/philosophical view underpins it. An imprecise formalization can lead to many unintended errors in an application. On the other hand, within for example information retrieval, the ontology typically serves as a background ontology supportive for relating different concepts and identifying mutual similarities. Thus, an ontology does not need to be 100 percent correct or coherent, but can be a loosely constructed graph without logical inheritance.

In reference to figure 2.1 A), an ontology appropriate for information retrieval could be any of those stated, whereas in tasks requiring more accuracy, an ontology should fulfill the requirements from the outer right end. Additionally, one should consider whether it is important to work with general ontologies like WordNet [94] or Cyc [67], or whether a domain ontology (presented in section 2.3.2), or a combination of both, is more appropriate to the task.

This chapter focuses on the concept of ontology. Ontology, as it is used in philosophy and computer science, is presented, along with some scientific philosophical methods for ontology modeling. Additionally, different levels of ontology are presented, such as top-level ontologies, general ontologies,

and domain ontologies.

The next chapter will take a closer look at formal KR and focus on the field of ontology within computer science and philosophy.

2.1 Ontology in philosophy and logics

Ontology concerns the nature of being, i.e. what exists in the world. The word ontology stems from ancient Greek and Parmenides and Aristotle have especially contributed key principles, syllogisms (reasoning rules), differentiae and categories [123, 52].

Later, these categories were ordered into hierarchical trees by, for example, Porphyry in the third century A.D. and Peter of Spain (later Pope John) in the thirteenth century A.D. who also introduced supposition and composition [123].

In the eighteenth century, the German philosopher Kant challenged the Aristotelian system, trying to provide a framework for classifying categories into groups of three that later resulted in the notion of thesis, antithesis and synthesis [123].

As a more formal treatment of the categories, in the nineteenth century, the mathematician Georg Cantor developed set theory from Boolean algebra, introducing individuals and the sets that individuals can belong to [70, 123]. Mathematically speaking, sets are structures collecting entities that can be manipulated using a membership-operation and the subset-operation. Some philosophical aspects of set theory state that sets can only be used to describe discrete things like dogs, integers and coins [123].[1]

Although Cantor might not be considered a founding father of ontology on the same level as Kant and Aristotle, his contributions to set theory are important for many people working in fields of ontology; in this thesis especially in section 3.1.3, which introduces class relationship logic.

Often, ontology within philosophy is referred to as formal ontology following Husserl, not to be confused with formalism for representing ontologies as presented in chapter 3, which is rather in line with the definition of e.g. Cocchiarella, as "the systematic, formal, axiomatic development of the logic of all forms and modes of being", from [57].

[1] This will be discussed and challenged in section 4.3

2.2 Ontologies in computer science

In philosophy, the concept of ontology concerns what exists and is a well-known discipline starting with Aristotle's Metaphysics [11, 124]. However, in computer science, the term has more varied definitions. Ontologies can range from a glossary (a list of words and their corresponding meanings) or a simple graph, to advanced systems that attempt to reflect the real world using concepts, individuals, relations, and constraints, etc. [104]. This variety and complexity is reflected in figure 2.1, in which different types of ontologies are placed on a complexity scale.

Philosophical ontologies will henceforth be referred to with a capital "O" in this dissertation and usage of the term within other semantic fields (typically in computer science) will be referred to with a lower case "o."

Gruber [55] describes an ontology very generally in the following way: "An ontology specifies a vocabulary with which to make assertions, which may be inputs or outputs of knowledge agents (such as a software program) (...) an ontology must be formulated in some representation language (...)."

This definition is highly oriented towards ontology in computer science and KR and also captures most of the ontological types in the figure 2.1 A) on the right hand side of the separating line. The definition is useful in reference to this dissertation, since most of the effort in representing knowledge has the purpose of creating assertions on the knowledge within the ontology.

In computer science, and especially artificial intelligence (AI), an ontology most often refers to an engineered artifact and consists of a vocabulary of concepts and relations as well as a set of intended meanings that make the assumptions explicit [56]. These artifacts can either be conceptualizations, pure formal systems or attempts to model the real world as precisely as possible.

When endeavoring to formalize an ontology, the type of formalisms used is of importance, since it can have impact on both reuse and integration with other, related ontologies and axioms [120, 56]. Section 2.2.1 will consider different methodologies for constructing ontologies, for example, within AI.

The terms *knowledge base* and *ontology* are often confused within computer science. Following Sowa, [123], a knowledge base is "rather an informal term for information in the collective form, including one or more ontologies, declarative or procedural rules concerning the knowledge." Thus, in the definition of Sowa, many task-dependent ontology systems are more appropriately considered knowledge bases, which contain more information than ontologies.

2.2.1 Conceptualism, realism and pragmatic implementalism

Contemporary treatments widely discuss the scientific method selected in a specific ontology modeling. Since ontology in itself is a meta-science, the modeling method will always be an important subject. Three views central to the debate on approaches to modeling ontologies are presented here.

This categorization is an attempt to clarify some choices made within this thesis, based mainly on the formulation of Guarino and Giaretta [57] and Smith, as presented in chapters one, four and five in [120].

A. Pragmatic implementalism: First, a task-dependent view in modeling approaches is described. This understanding, which will be called "pragmatic implementalism," is familiar to what is sometimes denoted as formalism or nominalism in a modeling context. Guarino calls this view *transfer-view* and it strictly focuses on the functional means for ascribing a certain goal, since whether objects exists or not is not important. It is one of the first ways in which the AI field converted knowledge into knowledge-based systems, based on mimicry of the expert's brain [58].

In this understanding, the system and the usage of the system are what should be predominant; the philosophical details within the ontologies such as properties, abstract objects and universals, or whether those exists in the real world, are not of interest [114]. An ontology or knowledge base should be designed to be useful for the task of the user, without a focus on the intensional meanings of the concepts or classes. Extensions of these concepts are, on the other hand, more important for this pragmatic view than other views, since the way agents speak and write about a subject is useful in many applications.

It is not contentious to this view to represent the same conceptualization in two different ways in different implementations. A famous example within ontology modeling is the decision of whether a type of "unicorn" can be a part of an ontology. Within this pragmatic ontological understanding, the answer is clear. If it is important for the task, a unicorn should be introduced as a class or object, depending on the system performance and other pragmatic issues.

B. Conceptualism: Conceptualists are often described as believing in formal, universal concepts, states of mind or mental objects that do not exist independently of human consciousness [20]. Objects only have a meaning in terms of how they are conceptualized. For example, a scientific concept is understood by a domain expert and exists as a kind of collaboration or agreement within the society of other experts. This methodology is covered by [57] which demonstrated ontology as "a particular conceptual framework at the knowledge level," as opposed to the view on ontology as an artifact

on the symbolic level (as in pragmatic implementalism).

The International Organization of Standardization (ISO) claims to have adopted a conceptual view on terminology work (ISO:704, Terminology work: Principles and methods [1]) by Smith [120], which he argues is problematic with respect to many uses, such as ontological merging. Guarino has criticized the view of conceptualism in its old form in e.g. [57, 58].

To a conceptualist, an ontology cannot have two different representations. If the same concept has two different meanings, it is part of two different ontologies and it is really two different concepts although they share the same name. Additionally, it is not important if a concept exist in the real world since the important issue is whether it exists as a mental object or not.

As a reflection of this, a conceptualist would answer "yes" to the question of whether a unicorn would be a part of an ontology, since there is a common understanding of what a unicorn is as a mental concept as well as of the properties it has.

C. Realism: In the contemporary ontology debate, Smith, [120] and Guarino [58, 56] both claim that ontologies should represent "reality," or the real world.

For Smith, (chapter 5 in [120]), realism is the central philosophical discipline for creating applied ontologies. This view facilitates the possibility of merging similar ontologies from two different domains, if they both attempt to model the same aspect of the real, underlying world from two perspectives. In biomedicine it is suggested that one attempts to model the laboratory context.

Guarino [56] advocates intensional relations, so-called conceptual relations, as reflecting reality. An ontology, in this view, is language-dependent, while a conceptualization is language independent. This reflects the view that an ontology will describe the world, which is possible using at-hand vocabulary. Conceptual units are mental constructs and, thus, should not be language-dependent, but may possibly be domain-dependent.

For a realist, discussing how many representations can be made of an ontology has no meaning. An ontology reflects the real world as it exists in e.g. the laboratory and, thus, it is always possible for a true realist ontology to be merged with other ontologies in classes that are equal or intersecting.

A realist would not consider unicorns as part of an ontology. They might be concepts in our heads, but they are not in the real world and thus cannot be part of an ontology. This also indicates that realism attempts to move towards Aristotle's conception of Ontology.

2.2.2 Scientific methodic approach in this thesis

All of the three above views on ontology are open to criticism in comparison to the direction and the choices made in this dissertation.

Although pragmatic implementalism seems to be a useful and tractable approach to introduce in modeling databases or to make a knowledge base work quickly and properly, it also has huge issues for modeling ontologies. For example, the system cannot be used in many contexts, because it will easily give erroneous results when introducing new reasoning rules. Another example is that mapping to other ontologies will be highly problematic.

On one hand, intensional analysis provides an abstraction that is useful for mapping to other ontologies and knowledge bases. On the other hand, an extensional description can be useful for disciplines as semantic annotation of domain information. This helps expert and non-expert users to retrieve information from, for example, different and scattered texts and knowledge bases. Thus, a deep analysis of both points of view offers a nuanced understanding of the concepts and relations within an ontology as well as the possibilities for using this knowledge in the more extensional parts of the world.

Whereas both realism and conceptualism have an intensional view on Ontology, they differ slightly in the focus on trying to mimic the real world and trying to mimic what we think of the real world.

Realism has its usefulness when building large, merge-able ontologies for several kinds of tasks performed on the ontology. However, realism is a high ideal that can be difficult to attain and a realism ontology will change every time new tools emerge as new aspects of the real world, e.g. new nanotechnologies that can build and track new properties concerning already-known materials.

Conceptualism has its optimal usage within education, knowledge acquisition and clarification of rather abstract concepts to understand, for example, a new field. This dissertation presents analysis on both the extensions and intensions of the regulatory relations. For example, a concordance study and a frequency study of verbs is performed, representing the regulation relations (presented in chapter 4), as well as an intensional semantic analysis of how these relations can be understood (from the viewpoint of realism).

Different methods for different approaches

While working on this thesis, it emerged that the scope of the task of ontology is important as a basis for deciding which of the scientific methods should be followed. In the foundational work of defining the semantics of the regulates relation, a realism approach was useful for the formal analysis of, e.g. granularity and relation types (section 4.4).

One could argue, for example, that the reasoning rules presented in sec-

tion 4.5 are purely rules of pragmatic implementalism not connected to the ontological understandings of any class, since they only approximate events in reality. However, such rules are experiments on the ontology and not axiomatic parts of the ontology. Thus, although developed prior to later work on the intensional and extensional semantics of relations, they should be seen in context, as a first step, for an information system built upon a realism ontology.

Finally, in the linguistic domain modeling in section 5.4.3, a top-ontology for defining the semantic types mapped into text is required. In this context, the pragmatic Unified Medical Language System (UMLS) top-level ontology Semantic Network had classes useful for the semantic roles of interest, not the idealistic, realist top ontology Basic Formal Ontology (BFO). In the work on meta-terms characterizing enzyme-inhibition in chemistry terminology, the method is more focused on concepts as understood by a group of experts than on how we reflect the real world (since the purpose is on coherence in concepts and education).

Thus applying one single methodology is not appropriate. Instead, approaches that seem relevant to this dissertation, whether educational, modeling or in implementations, will be employed, e.g. semantic tagging.

2.3 Types of ontologies

Just like any other domain, the domain of knowledge base ontologies requires some disambiguation for clarification. In the literature, there is plethora of different kinds of ontologies, differing not just in philosophical view (such as realism versus conceptualism), but also with respect to other differentiae like purpose and level of granularity.

In figure 2.1, different ontologies are ordered with respect to their richness or expressiveness. The ontologies on the right-hand side are those typically used within AI as well as for reasoning over the ontology.

Most common is the distinction between top ontologies, domain ontologies, and general ontologies, which differ with respect to level and whether they are domain specific of for general linguistic purposes.

While top ontologies attempt to capture the most basic concepts and provide a basis for mapping other ontologies, the design of these has mainly been the domain of philosophers. From Aristotle's categories to BFO and DOLCE [11, 35, 49], the speculations and developments have been of a very foundational and philosophical nature.

General ontologies like Cyc [67] and WordNet [94] are mostly used for broad, non-domain specific purposes and developed by linguists as vast common language repositories for glossaries, general information retrieval, and comprehensive lexical information.[2]

For the most part, information scientists who work within the field or domain they describe, design domain ontologies. Examples of domain ontologies are the resources of the UMLS [24], where SNOMED CT [72] is mainly in the medical domain, and Gene Ontology [14] focuses on the domain of molecular biology, for example. Domain ontologies are usually attached to some tasks within the field that they describe. Finally, a collection of both top-ontology and domain ontologies can be referred to as a universal ontology.

2.3.1 Top Ontologies

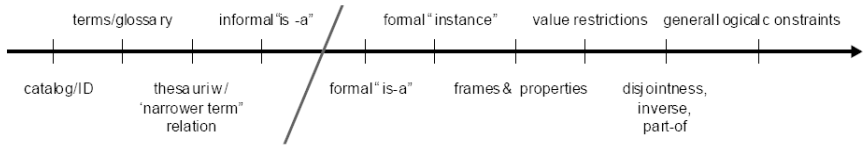
As is presented in section 2.3, the work on developing top ontologies has traditionally been a discipline that appealed to philosophers.

Only a few top ontologies will be presented here, namely Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), Basic Formal Ontology (BFO), Suggested Upper Merged Ontology (SUMO) and Semantic Network. Of these, Semantic Network differ the most as a domain top ontology attached to the aforementioned UMLS. Many more exist and are, for example, discussed and compared in [52] and chapter eight in [120].

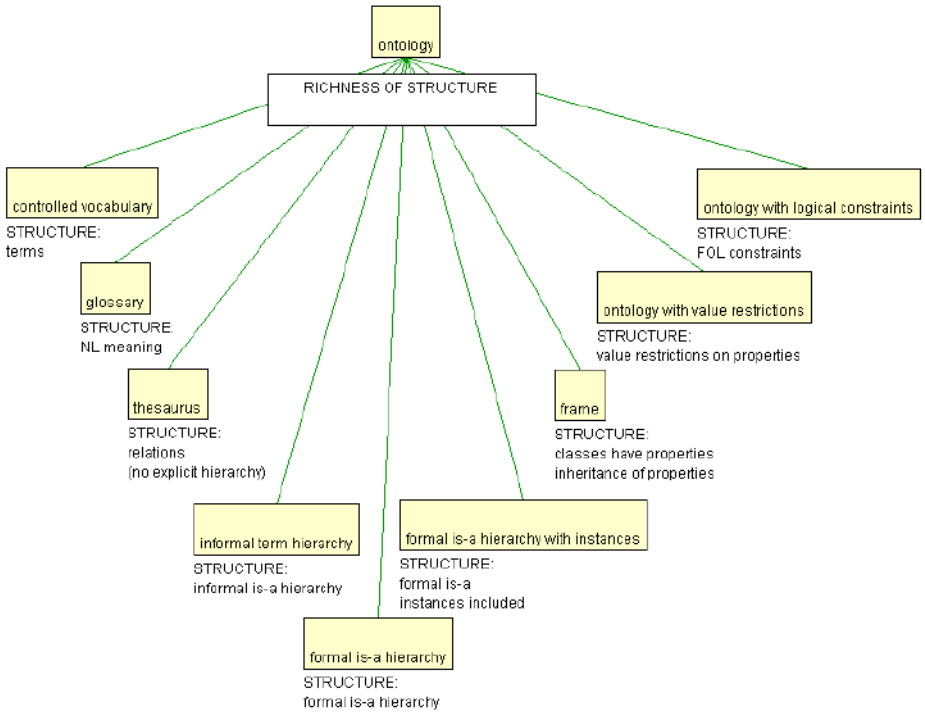
[2] These will not be discussed further in this section since focus has been on domain ontologies and top-ontologies.

Figure 2.1: An ontology's level of complexity grows with the amount of meaning that it is possible to express. A) concerns the richness of the structure [104]; B) identifies *values* of the structures [86].

A)



B)



General, top-level ontologies include DOLCE [49], BFO [35, 122] and SUMO[125]. Both DOLCE and BFO have a philosophical view and a formalization in first-order logic (section 3.1.1). DOLCE has a quite modest vision:

> *(...) we do not intend DOLCE as a candidate for a "universal" standard ontology. Rather, it is intended to act as starting point for comparing and elucidating the relationships with other future modules of the library, and also for clarifying the hidden assumptions underlying existing ontologies or linguistic resources such as WordNet. As reflected by its acronym, DOLCE has a clear cognitive bias, in the sense that it aims at capturing the ontological categories underlying natural language and human commonsense(...) [49]*

BFO [35, 122], on the other hand, claims to be more task oriented:

> *(...) it is narrowly focused on the task of providing a genuine upper ontology which can be used in support of domain ontologies developed for scientific research, as for example in biomedicine within the framework of the OBO Foundry. Thus BFO does not contain physical, chemical, biological or other terms which would properly fall within the special sciences domains [35].*

SUMO [125] is even more focused on the usage in a pragmatic implementalist approach, since it is the only top ontology mapped to all of WordNet:

> *The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today. They are being used for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has been mapped to all of the WordNet lexicon.* [3]

The Semantic Network of UMLS [92] is, contrary to the three previous top ontologies, domain specific. It is not formalized in logics like DOLCE and BFO; however, every concept and term in the resources of UMLS is linked to a concept in the Semantic Network. A special feature concerning this ontology is that it is built to link resources together rather than built based on high-level abstractions and formalization methods.

This makes the Semantic Network adequate for e.g. usage in semantic annotation as will be shown and discussed in section 5.3. The Semantic Network and its usefulness in semantic annotation for knowledge retrieval will be further described in section 5.4.4.

[3] *www.ontologyportal.org/, Adam Pease, 2011*

2.3.2 Domain ontologies

For a task that uses a specific field of knowledge, it can be more adequate to utilize domain ontologies in addition to, or instead of, the more general ones. A domain ontology, which is here understood as an ontology that is applied to a certain domain, is either very narrow, such as the MGED Ontology - an ontology for microarray experiments [129], or more general, such as the medical clinical SNOMED CT for the domain of health care [72].

In the ontological field, domain ontologies already exist, and initiatives like OBO [118] and UMLS [24] attempt to combine and collect common, area-specific ontologies.

N. Guarino classifies ontologies and has not only provided a definition of domain ontology, but also proposed which tasks it can be used for. The ontology is, according to Guarino, the domain of discourse, i.e. "Any formal theory is a theory about a domain" [56]. To Guarino, a terminological ontology is defined as a domain ontology, which will be exemplified in section 5.1.

While philosophers primarily develop top ontologies, domain ontologies often have a broader spectrum of developers, such as terminologists, working in cooperation with domain experts (as is the case for enzyme inhibition in this dissertation [38] and in the SIABO-collaboration [9]). Essentially, creators from many fields must carry out a domain ontology, in order to describe the domain of discourse.

Lately, approaches toward automatically and semi-automatically extracted domain ontologies from domain literature have been developed [90]. In a sense, the linguistic work carried out in chapter 5 can be seen as a product of this tendency, utilizing the semantic frames and knowledge patterns developed within these sections. is a product of this tendency and utilizes the semantic frames and knowledge patterns developed within these sections. Domain ontologies developed with this purpose often reflect an extensional ontology modeling extracting *how things are described* in natural language rather than the more intensional, which models *how things are meant*.

2.4 Summary

This chapter introduced the meaning of ontology vs. Ontology (within philosophy), presenting some of the main scientific methods in ontology research in computer science, namely realism, conceptualism and what is termed here as pragmatic implementalism. The use of these methods in the different works upon which this dissertation is based was discussed and compromises in relation to usage of the resulting ontology delineated. Furthermore, different levels of ontologies were introduced: top level, general and domain ontologies.

Chapter 3

Knowledge representation I: Formal foundation

A computational knowledge representation(KR) needs a representative language. In a broad understanding, depending on the usage, this level of ontology could be anything from natural language to purely symbolic representation.

To choose the right formalism is both a question of its tractability within information systems as well as its expressiveness with respect to its domain (i.e. what do we want to get out of the knowledge? And, what role should the knowledge representation have, e.g. [39]). In an age where the efficiency of computers is ever-prevalent, the role of the knowledge representation is remarkably important.

Often, the choice of formalism is biased towards the following elements:

- Old routines. When people are used to representing information and knowledge in a relational database, they keep on using this device.
- "Me too" formalisms. Everybody else within the field uses the formalism (as for example OWL), thus one should use it as well - at least for comparison.
- "Expressiveness buffering." Use of formalism with the option of expressing more than apparently needed, if more expressiveness is needed in a later implementation.

Although these points are expressed rather ironically, it is not postulated that any of these reasons are bad foundations for deciding a formalism, as long as other motivations for supporting those are possible. For example, the implementation can be faster if an old routine is followed, although it might not capture everything. Likewise, if the ontology should be merged with other ontologies, it makes sense to use the existing format.

However, if everything can be expressed in a large system on the basis of simple relational algebra, it would be appropriate for the majority of ontologies for their concepts only to be defined based on their relations to other concepts. This representation corresponds to an (advanced) graph-form similar to the knowledge structures that computers normally handle.

In description logic($\mathcal{DL}$), using e.g. OWL, more advanced descriptions are often created to attribute the concepts within a concept system. Additionally, this expression of knowledge in a more advanced format than needed will be referred to as "expressiveness buffering." It can be useful if the future task of the KR is uncertain, but it can also complicate the ontology unnecessarily with a diminished tractability as a result.

Chapter 4 will briefly present the existing representations of the domain of regulations in biomedical knowledge. For this, an understanding of the formal languages available for representation of knowledge is needed.

This chapter will elaborate on formalisms - and in particular logic-founded formalisms - that can be utilized to represent domain knowledge. Often, biomedical knowledge is represented in logics for the purpose of classifications and reasoning.

Additionally, work on reasoning will be presented, utilizing the complex role-inclusions published in [4]. In the discussion (section 7.3.1), the advantages of these role-inclusions are argued for using an example in the Gene Ontology.

3.0.1 Notation

Different notations that appear in various published papers have been revised and built upon in this thesis in order to unify them. The notations are generally in line with typical formal ontology notation such as description logic, and designed to make a clear distinction between individuals and classes as well as individual relations and class relations.

Lowercase letters "$x, y, z \ldots$" are used as variables ranging over arbitrary individuals (or tokens, elements, instances, or particulars) and capital letters "$C, C_1, C_2 \ldots$" or "$A, B, C \ldots$" are used as variables to range over classes (or concepts, sorts, kinds, or types).

Relations among individuals are in bold face, with a lowercase first letter, e.g. "x **rel** y" or "x **hasPart** y" whereas relations among classes are italicized, with a lower case first letter e.g. "C_1 *rel* C_2" or "C_1 *hasPart* C_2". Particular relations are often called predicates, roles, attributes or slots.

Table 3.1: Notations.

Name	**Description**	**Notation**
Individuals	(tokens, elements, instances or particulars) - Entities belonging to a certain class.	$x, y, z \ldots$
Classes	(concepts, sorts, universals, kinds or types) - Ontological entities containing (up to) several individuals.	$A, B, C, C_1 \ldots$
Individual relationships	Relations among individuals	x **rel** y
Class relationships	Relations among classes	C_1 *rel* C_2
Roles	(predicates, attributes, slots) - Particular relations such as semantic annotations to elements in a text	AGT:, PTN:
Constrained relations	Conceptual relation with constrained relata types, e.g. if a and b are ontological type restrictions and *rel* is a relation $< a - rel - b >$.	rel_{ab} or $[rel,\ AGT : a,\ PTN : b]$
Part of speech (POS) tagging	Grammatical tags based on syntax of the tagged sentence based on Penn Treebank	\VB
Relation operator	Operator that provides possibility for extracting e.g. the continuant from a process, such as *glucose* $\mathbf{stimulates_{cp}}$ $production\_of(insulin)$	c_1 $\mathbf{stimulates_{cp}}$ $production\_of(c_2)$

3.1 Logic-based formalisms

This section introduces four different kinds of formalisms that all have some shared features. In this context, a formalism is a language based on mathematical principles, which can transform a knowledge subject in calculus, reasoning, and axiomization, in line with the vision of Leibniz [83].

In this sense, both first-order logic, functional algebra and SQL, as well as the concept in algebra of ONTOLOG [100] are formalisms in which knowledge can be formalized such that it is subject to operations.

I have chosen to mainly explore logic for knowledge representation in the formalisms first-order logic($\mathcal{FOL}$) and the $\mathcal{FOL}$-subset class-relationship logic($\mathcal{CRL}$) and the $\mathcal{CRL}$-part of description logics($\mathcal{DL}$). An additional section on concept algebra and the language ONTOLOG [100] is included, since the notation of this is used in section 5 and 6.

All of these formalisms are rather expressive, which is useful for the purpose in defining the deeper semantics of regulation. Whether applications should also be based on these expressive logics will be discussed in chapter 6.

3.1.1 First-order logic($\mathcal{FOL}$)

First-order logic($\mathcal{FOL}$) was developed in the nineteenth century, primarily by Peirce [107, 123] and Frege [48] and is based on the more simple propositional logic and Boolean logic introducing the logical quantifiers $\forall$ and $\exists$ [123]. First-order logic is used to reason about individuals and their properties, and allows for quantification of these individuals, whereas second-order logic can be used to reason over classes of classes or predicates.

$\mathcal{FOL}$ can be considered the most fundamental logic by which most/all other logics can be described; first-order logic has a relatively simple reading and generality. Thus, some of the later formal relations represented have a direct translation in $\mathcal{FOL}$ and many fragments of the logic can be implemented in a decidable, tractable way.

The language of first-order logic is built from propositional connectives such as "and" ($\wedge$), "or" ($\vee$), conditionals ($\rightarrow, \leftrightarrow$), "not" ($\neg$), predicate symbols involved in e.g. $P(x)$ or $Q(x, y)$, variables $x, y, z, \ldots$, and quantifiers $\forall$ and $\exists$ (reading "for all" and "there exists").

First-order logic does not come with a built-in potential for gaining the expressiveness of differential equations as those presented in section 4.2.3. But quantifiers $\forall, \exists$ ddetermine whether the predicate involves all individuals, or whether we know that at least one element exists. The approach gains tractability by not dealing with numeric quantifications, however, on the other hand, it has higher complexity when dealing with the quantifiers and negations compared to simple propositional logic or relational algebra. This will be elaborated in the discussion in chapter 7.2.

3.1.2 Description logics ($\mathcal{DL}$)

Description logics($\mathcal{DL}$) is a widely used formalism for representing knowledge, especially within the community of biomedical ontology research. It is a decidable fragment of $\mathcal{FOL}$ and has been developed since the 1980s, with a tight connection between theoretical research and practical implementation of systems [15].

Description logics is often considered one logic language, even though it is a family of languages with different levels of expression. Although the objective of creating this formalism has been to create decidable logic systems, it is possible to construct a $\mathcal{DL}$ that is undecidable, or slightly intractable.

SNOMED CT uses a light version of $\mathcal{DL}$ called $\mathcal{EL}+$, consisting of existential quantifier, conjunction, composition of relations and top element, $\{\exists R.C, \sqcap, \circ, \top\}$ [72].

A central feature within $\mathcal{DL}$ is its division into a T-box, which contains concept (or class) definitions (or terminological definitions) and an A-box which contains assertional knowledge on individuals rather than concepts (or classes). For example a T-box-fragment from the Gene Regulation Ontology [21] in $\mathcal{DL}$:

$$\begin{gathered} NegativeRegulation \equiv \\ RegulatoryProcess \sqcap \exists hasQuality.NegativePolarity, \end{gathered} \tag{3.1}$$

corresponding to the $\mathcal{FOL}$-sentence:

$$\begin{gathered} \forall x(NegativeRegulation(x) \leftrightarrow \\ (RegulatoryProcess(x) \wedge \exists y\,(NegativePolarity(y) \wedge x\,\mathbf{hasQuality}\,y))) \end{gathered} \tag{3.2}$$

An A-box example meaning "experiment_233" is an individual of the class "inhibition of glucose transport" can be written:

$$InhibitionOfGlucoseTransport(EXPERIMENT\_233). \tag{3.3}$$

If we want the assertions to be more general, showing overall negative regulation, the expression would be:

$$NegativeRegulation(EXPERIMENT\_233). \tag{3.4}$$

Finally, an individual can demonstrate an instance of negative regulation, such that "inhibition of glucose transport" in itself is an individual, although this does not seem like a natural individual:

$$NegativeRegulation(INHIBITION\_OF\_GLUCOSE\_TRANSPORT). \tag{3.5}$$

The elements of the T-box overlap with $\mathcal{CRL}$, presented in the next section, except for the open world assumption existing within the $\mathcal{DL}$ framework and not in $\mathcal{CRL}$/logic computer languages such as Prolog.

As also mentioned above in section 3.1.3 and displayed in table 3.2, class relationships from the popular knowledge representation language $\mathcal{DL}$ [15] is defined by its relation among individuals as:

$$C_1 \sqsubseteq \exists rel.C_2 \quad \text{iff} \quad C_1\ rel_{\forall\exists}\ C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \exists y(x\ \mathbf{rel}\ y \wedge C_2(y))). \tag{3.6}$$

Additionally, the $C_1\ rel_{\forall_{only}}\ C_2$ relationship corresponds to the value restriction $\forall R.C$ in $\mathcal{DL}$ as seen in the table 3.2.

3.1.3 Class relationship logic

In knowledge based systems, class relationships have played a bigger role in the representation of knowledge. For example, in biomedical ontologies the OBO foundry operates with relations among classes based on the relations among the individuals in these classes [121, 119]. In e.g. [22], a framework is presented on the formal and informal properties of class-relations as well as other ontological relations within a semantic web-approach to, for instance, biomedical ontologies.

The reason for using relations among classes,[1] rather than relations among individuals, is that classes, if seen as reflecting the real world, are really the main focus in many ontologies such as biomedical ontologies, in which individuals are often considered empirical data. The terminology with which classes or concepts are discussed reflects an attempt to abstract from individual to class when modeling ontologies.

Additionally, the rationale for reasoning over class relations is that the approach can be useful within drug discovery as noted in e.g. [112]. For example, the link between a drug and a disease might be deduced by a chain of events, which can be formed logically as compositions of (different) class relations.

In the ontology formalism, description logic [15] the universal and existential restrictions are, by the definition given in this dissertation, also relations among classes.

Class relationships are relations based on $\mathcal{FOL}$, where individuals always have a membership in a class and there is a closed world assumption [102].

A relationship among classes $C_1\ rel\ C_2$ is based on relations among individuals but with necessary quantifiers refering to individuals, such as:

$$\forall x(C_1(x) \rightarrow \exists y(x\ \mathbf{rel}\ y \wedge C_2(y))). \tag{3.7}$$

[1] Note that we use *relation between classes* and *class-relations* interchangeably although the latter is not strictly mathematically correct without a second order definition.

The usefulness of this representation/formalism is found in binary relations among classes or sets, rather than binary relations among individuals. The notation, $\mathcal{CRL}$, was developed by J. Fischer Nilsson and material on this was published recently in [102, 101].

The motivation for a logic on classes was presented recently in, for example, [119, 23, 121, 22] and is connected to the insight that, in biomedical ontologies, knowledge is often expressed in terms of classes or types rather than actual individuals. It is indirectly used in the T-boxes in description logics, for example as is described in equation (3.2).

A binary class relationship as that of, e.g. description logics [15] and the more general form of class relations presented in [121, 119], is defined by its relation among individuals, such as:

$$C_1 \; rel_{\forall\exists} \; C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \exists y(x \; \mathbf{rel} \; y \wedge C_2(y))). \tag{3.8}$$

In this case it is a $\forall\exists$-relationship, which reads "forall-exist relationship." Table 3.2 provides an overview of the nine different binary combinations of quantifiers[2] and their relation to $\mathcal{FOL}$ and $\mathcal{DL}$. All $\mathcal{CRL}$-relations can be formulated in $\mathcal{FOL}$. When applied to taxonomies, logical inferences of the relations will be generated if implemented correctly.

A list of inference rules for this logic can be found in a basic form, easily implemented in Prolog sublanguage Datalog (exemplified in [102]). The first four relationships in table 3.2 can easily be constructed in Datalog. It is possible to add extra reasoning rules if required, for example, based on domain and the tasks the system should perform.[3]

Additional work on $\mathcal{CRL}$-compositions is presented in section 3.2 [4] and exemplified in the semantic analysis in section 4.4 [136].

Furthermore, some of the relations have a direct parallel to $\mathcal{DL}$ whereas others have more spectacular descriptions. Note that the relationships $r_{\exists\forall}$ and $r_{\exists\forall only}$ might have correspondences in $\mathcal{DL}$, though these have not been identified yet.

$\mathcal{DL}$ and "$rel_{\forall\forall}$"

Table 3.2 shows the two class relations $rel_{\forall\exists}$ and $rel_{\forall only}$(defined in (4.6) and (4.7)) can be formalized in $\mathcal{DL}$ by:

$$\begin{aligned} C_1 \; rel_{\forall\exists} \; C_2 \quad &\text{iff} \quad C_1 \sqsubseteq \exists \mathbf{rel}.C_2, \\ C_1 \; rel_{\forall only} \; C_2 \quad &\text{iff} \quad C_1 \sqsubseteq \forall \mathbf{rel}.C_2. \end{aligned}$$

[2] These can be collapsed to five that can describe all if we introduce the inverse relation rel^{-}.

[3] The class-relationships containing the "only" constraint will need an integrity constraint in the implementation, and this has not been developed yet in the framework in [102]. However, the promising results in [77] hint that at least classification and instance cheking can be carried out in Datalog.

Table 3.2: The different class relation-types in first-order logic, description logic (when possible) and natural language [136, 4, 102]. Only the first four relationships are easily implementable in Datalog.

$\mathcal{CRL}$ $(A\, r\, B)$	$\mathcal{FOL}$ formulation	$\mathcal{DL}$	Natural language formulation
$r_{\exists\exists}$	$\exists x\big(A(x) \wedge \exists y\big(x\ \mathbf{r}\ y \wedge B(y)\big)\big)$	$A \sqcap \exists rel.B \nvdash \bot$ or $A \sqcap \exists rel.B$ *	There exists an element in A that is related to some element in B
$r_{\forall\exists}$	$\forall x\big(A(x) \rightarrow \exists y\big(x\ \mathbf{r}\ y \wedge B(y)\big)\big)$	$A \sqsubseteq \exists rel.B$	Every element in A is related to some element in B
$r_{\forall\forall}$	$\forall x\big(A(x) \rightarrow \forall y\big(B(y) \rightarrow x\ \mathbf{r}\ y\big)\big)$	$A \sqsubseteq \forall(\neg rel).\neg B$	Every element in A is related to every element in B
$r_{\exists\forall}$	$\exists x\big(A(x) \wedge \forall y\big(B(y) \rightarrow x\ \mathbf{r}\ y\big)\big)$		There exists an element in A that is related to every element in B
$r_{\forall\forall o}$	$\forall x\big(A(x) \rightarrow \forall y\big(x\ \mathbf{r}\ y \rightarrow B(y)\big)\big)$	$A \sqsubseteq \forall rel.B$	Any element in A is *only* related to elements inside B
$r_{\exists\forall o}$	$\exists x\big(A(x) \wedge \forall y\big(x\ \mathbf{r}\ y \rightarrow B(y)\big)\big)$		There exists an element in A that is *only* related to elements inside B
$r_{\forall_o\exists}$	$\exists y\big(B(y) \wedge \forall x\big(x\ \mathbf{r}\ y \rightarrow A(x)\big)\big)$		There exists an element in B that no element outside A is related to
$r_{\forall_o\forall}$	$\forall y\big(B(y) \rightarrow \forall x\big(x\ \mathbf{r}\ y \rightarrow A(x)\big)\big)$	$B \sqsubseteq (\forall rel^-).A?$	There are no element outside A that is related to any element in B
$r_{\forall_o\forall_o}$	$\forall x\forall y\big(x\ \mathbf{r}\ y \rightarrow \big(B(y) \wedge A(x)\big)\big)$	$B \sqsubseteq (\forall rel^-).A \sqcap$ $A \sqsubseteq \forall rel.B$	It is *only* elements in A and B that are related

*If we consider the classes non-empty.

Other class relations such as $rel_{\forall\forall}$ and $rel_{\exists\exists}$ cannot be expressed in a majority of description logics. However, in very expressible description logics including full concept negation and role negation [85, 84], for example, the $rel_{\forall\forall}$ relation can be formalized by:

$$C_1 \; rel_{\forall\forall} \; C_2 \quad \text{iff} \quad C_1 \sqsubseteq \forall(\neg\mathbf{rel}).\neg C_2. \tag{3.9}$$

Alternatively, a new operator in line with the $\exists\mathbf{rel}$ and $\forall\mathbf{rel}$ operators could be added to a description logic. Such an operator as in equation (3.9) has already been added to similar modal logics and goes under the name "the window operator". However, a minimal description logic with this operator has apparently not been investigated.

3.1.4 Concept algebra

So far in this section, the presentation of formalisms has focused on expressiveness and tractability, on whether the language is decidable and on the symbols it contains with potentially many possible applications.

Concept algebra, on the other hand, has a quite narrow scope in terms of a direct linguistic application. This application is to formalize concepts within a lattice-based algebra such that they represent calculable and tractable semantic structures, defined by: $L := \{\wedge, \vee, \leq, \geq\}$,where the operators denotes: meet ($\wedge$), join ($\vee$) and conditional operators ($\leq, \geq$).

Thus, in its structure, concept algebra is simpler than the framework of $\mathcal{FOL}$ and many $\mathcal{DL}$-languages, though still able to represent a taxonomy, conjunction and disjunction.

Ontolog. The ontolog language is built upon concept algebra and aims to represent natural language as conceptual semantic structures. Within the framework of the language ontolog, semantic roles are seen as binary relations, and the resulting structures as *concept feature structures (CFS)*, which form the ontotypes/concepts recursively [100].

A CFS is defined as: $c\,[r_1 : c_1, r_2 : c_2, \ldots, r_n : c_n]$. In this structure, c is a concept (or ontotype), $r_1, r_2, \ldots, r_n$ are semantic roles such as "agent", "patient" or "source". $c_1, c_2, \ldots, c_n$ are CFSs (or concept arguments or values), which could be atomic, representing a simple concept like "insulin" or a complex structure like "insulin[Source: beta-cells]."

The structure is recursive and the relation between a semantic role and a concept feature structure, $r_1 : c_1$ is called a *feature-value* pair and has the form: $[r_1 : c_1, r_2 : c_2, \ldots, r_n : c_n]$. This functions as conceptual specialization of the head concept c. In other words, $c\,[r_1 : c_1]$ is always situated as a subtype of the node c in the ontology [9, 100]. Thus, it has an underlying lattice concept algebra.

In this way, new paths lead to more specialized concepts in the ontology, which is referred to as a generative ontology. The relations within ontolog

function as case roles, which are linguistic roles that add to lexical items [45]. In semantic indexing as described in section 6, a CFS reflects how concepts occur in text in line with the way linguists denote semantic roles [108].

Furthermore, to restrict which structures are admissible, the so-called ontological affinities, which restrict the type of relata (concepts around a relation or role) can be used. These are specified as triples $< c^1, r, c^2 >$ (which equals $rel_{c^1c^2}$ as used in [136, 138]), for restricting types and relations. The affinities can be specific, e.g. to rule out category mistakes. Ontological affinities are used in sections 4.4 and 5.3 in connection to the relations and relata connected with biomedical regulatory events.

In this thesis, role abbreviations like: $inhibition[PTN : glucose\_transport]$ are used as in [100] whereas ontological affinities of the relata types c^1 and c^2 are described by the notation: $rel_{c^1c^2}$. These relata types are synonymous with what is often called argument-types (of a predicate).

Generative ontology. A generative ontology is a form of ontology that can be built using an infinite-concept algebra (e.g. in ontolog) [9, 103].

A generative ontology is based on a finite ontology, with the concept-inclusion relation *isa* (this taxonomic structure is termed a "skeleton ontology"). An example of this is an inclusion path like: *insulin secretion* isa *secretion* isa *process* isa *event*.

A set of semantic roles can potentially expand the (taxonomy) ontology and contribute to the generativity of the ontology. Thus, a generative ontology is understood as an infinite set of concepts, thereby reflecting that phrase structures in natural language can be produced recursively. The name and concept are inspired by the notion of generative grammars and the generative lexicon by [110].

This kind of generativity makes it possible to map complex linguistic structures (like "insulin stimulates glucose transport") into correspondingly complex concepts (such as $disease[CausedBy : lack[WithRespectTo : insulin]]$), associated with nodes in the ontology. Additionally, this generativity corresponds to linguistic forms found in a text or query, such as "diseases caused by insulin lack," "diseases induced by insulin deficiency," "insulin deficiency disease," etc.

An application of a generative ontology is to collect text phrases and denote these with semantic information, for example, using CFSs as presented above.

3.2 Compositions of $\mathcal{CRL}$

This dissertation reduces the potential class-relation composites to create a logical foundation for complex role inclusions on class relationships introduced in section 3.1.3. $\mathcal{CRL}$ is based on relationships among individuals belonging to these classes of interest. These role inclusions can both be based on logical axioms of inheritance or reasoning rules/relational closures.

A total of nine possible different binary class relationships are considered, namely those in table 3.2, which is partly based on [102]. The class relationships and their motivation will be introduced in section 3.2.1, and the degree of knowledge that the different class relationships represent are discussed afterwards, adding examples from biomedicine.

3.2.1 Class relationships and role inclusions

Recall that in this dissertation, *class* refers to what is often called concepts, types or kinds, representing reality, which is mainly in line with e.g. [119].

With regards to individuals, complex role inclusions based on compositions are well studied and formalized, e.g. in description logics ($\mathcal{DL}$) [15] where they can be modeled as reasoning rules in e.g. the $\mathcal{DL}$-language $\mathcal{EL}+$ [16] using the adopted notation: $\mathbf{r}^3 \sqsubseteq \mathbf{r}^1 \circ \mathbf{r}^2$ for relations among individuals.

Introducing a $\odot$-connective, our subject of investigation is thus combinations of the form $A\,(r^1_{\forall\exists} \odot r^2_{\forall\forall})\,C$, for $A\,r^1_{\forall\exists}\,B$ and $B\,r^2_{\forall\forall}\,C$, which for the given example is defined as:

$$\begin{aligned} A\left(r^1_{\forall\exists} \odot r^2_{\forall\forall}\right) C &\equiv A\, r^1_{\forall\exists}\, B \odot B\, r^2_{\forall\forall}\, C \\ &\equiv \forall x \left(A(x) \rightarrow \exists y \left(x\mathbf{r}^1 y \wedge B(y)\right)\right) \\ &\quad \wedge \forall y\big(B(y) \rightarrow \forall z\big(C(y) \rightarrow y\mathbf{r}^2 z\big)\big) \end{aligned} \tag{3.10}$$

Thus, the goal is to determine which type the resulting class relationships will have given the type of two relations in the composite, i.e. given any combination of the quantifiers specifying the type of the relations r^1 and r^2 between classes A and B, and B and C respectively (i.e. $\forall\exists$ and $\forall\forall$ in the example above). We will examine whether a class-relationship between A and C of any of the considered types in table 3.2 exists, and if so, which it is.

Notations and formal definitions will be described in section 3.2.1. Eighty-one combinations of class-relationships based on their individual relationships were examined. Table 3.3 provides an overview. Arguments for their correctness are described in [4], and the theorems in table 3.3 are also tested with the higher order theorem prover using the TPTP-platform [126] using LEO-II and TPS.

Degrees of knowledge The actual application of the reasoning rules in table 3.3 and definitions of meaningful inferences on a certain domain is a task of knowledge engineers. These modeled rules are based on domain specific knowledge and complement the pure logic inferences. Thus, it can be useful to consider the levels of abstractions that can be used within the domain.

Knowledge using class-relationship representation can come in many forms and are based on the *degree of knowledge* your knowledge base contains. Some examples are discussed as follows:

The $\exists\exists$ relation represents the least degree of knowledge. In this case it is only known that an instance from the first class has a relation to an instance from the second class. This is not of much use if one is presented with a concrete instance of either of the classes; given an amount of insulin and a glucose transport one cannot infer anything about them from the knowledge that "$Insulin\ stimulates_{\exists\exists}\ Glucose\_transport$". An appearance of knowledge on this form could be a semantic extraction from, for instance, a biomedical text. The relationship has been detected, but nothing general should be concluded from just one (uncertain) example.

A further natural step in the degrees of knowledge may be the $\forall\exists$ relation. This knowledge might be obtained by observing a larger sample of instances from the first class. The knowledge "$Insulin\ stimulates_{\forall\exists}\ Glucose\_transport$" makes it possible to infer something about insulin, namely that any amount of insulin makes it possible to find a glucose transport that it stimulates. However, this knowledge does not really provide any information about glucose transport. This relation type is typically found in canonical ontologies and the subsumption-relation can also be seen as this relation type.

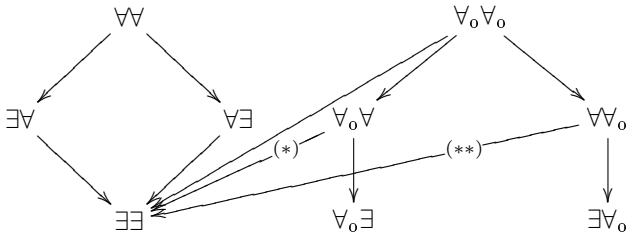
Continuing, the $\forall\forall_o$ relation will intuitively represent a higher degree of knowledge. But in the case of insulin and glucose transport it still only provides information about insulin, for example whatever insulin stimulates, it must be a glucose transport. But it does not tell us that we can always stimulate glucose transports with insulin. It is thus typically used to create constraints for the first class, e.g. "$Insulin\ isSecretedBy_{\forall\forall_o}\ Beta\_cells$", saying that whenever insulin is secreted, it is by a beta-cell.

Likewise, the inverse relation behaves. If r is $\forall\forall_o$ then r^{-1} is $\forall_o\forall$ and the statement above would be equal to "$Beta\_cells\ secretes_{\forall_o\forall}\ Insulin$".

The "final" degree of knowledge as represented by the $\forall\forall$ relation gives us information about both insulin and glucose transports. This is often the really valuable information, namely that given an arbitrary glucose transport we can always stimulate it with any given amount of insulin. This knowledge is typically found in interaction databases such as [73].[4] Notice, however,

[4]However, this does not prevent the $\forall\exists$ in being the underlying relationships in e.g. popular presentations of the knowledge in texts and speech in which people tend to generalize.

Figure 3.1: Ordering of the different relation types, assuming the relation is non-empty. $(*)=(M\times B)\cap r\neq\emptyset$,$(**)=(A\times M)\cap r\neq\emptyset$ [4]



that the logical strength of $\forall\forall$ is not necessarily larger than the one of $\forall\forall_o$; there is also information contained in knowing, that e.g. insulin *cannot* be secreted by anything *else* than beta cells. $\forall_o\forall_o$ is "stronger" since it restricts both the relata classes.

These considerations show that when specifying formal relations for knowledge representation in biomedicine, one should not only consider how the world actually is, i.e. what the ontology looks like, but also what kind of knowledge can be useful to represent in an application in the first place.

Mathematically speaking, the relations can be ordered according to their degree of knowledge following the diagram in figure 3.1, assuming that all concerned classes and relations are non-empty. This should be read so that if r e.g. is a $\forall\forall$-relation, then it is also a $\forall\exists$-relation as well as an $\exists\forall$-relation, and (due to transitivity) it is also an $\exists\exists$-relation, given that the relation is non-empty. This is easily seen by the (first order) definitions of the relationship types (see table 3.2).

As described above, a $\forall\forall$-relation is considered to be more informative than a $\forall\forall_o$-relation. Meanwhile, a $\forall\forall$-relation is not a $\forall\forall_o$-relation; though, under certain (natural) conditions described below, it is an $\exists\exists$-relation. In the example above with the $\forall\forall_o$-relation *stimulate* between *Insulin* and *Glucose_transport*, we are not told that there is not anything *else* besides insulin that could stimulate glucose transport. But if the (further) assumption is made that insulin indeed stimulates something at all, then it can be inferred that *stimulates* is an $\exists\exists$-relation. This requirement is that $(Insulin \times M) \cap stimulates \neq \emptyset$, where M denotes the entire domain.

This requirement corresponds to the intuition behind our knowledge representation, though this is often an implicit assumption; when we say that insulin stimulates glucose transport. Implicitly excluded is the case where something that is not insulin stimulates something that is not a glucose transport, while no insulin stimulate any glucose transport, even though this, mathematically speaking, would still qualify as a valid $stimulates_{\forall\forall_o}$-

relation between *Insulin* and *Glucose_transport*.

As a final remark to this diagram, one might think that if r is a non-empty $\exists\forall_o$-relation, then it would likewise be an $\exists\exists$-relation, but this is not the case as can be seen from the example $A = \{a, b\}$, $B = \{b\}$ and $r = \{(b, a)\}$. Unlike the example above, the extra restriction on r that would ensure it being an $\exists\exists$-relation would, in itself, imply that r is an $\exists\exists$-relation. This could also indicate that the usage of the $\exists\forall_o$-relation is limited within knowledge representation.

Definitions Notationwise, we write "$\exists x \in X(\ldots)$" for "$\exists x(X(x) \wedge \ldots)$" and "$\forall x \in X(\ldots)$" for "$\forall x(X(x) \rightarrow \ldots)$."
Formally, the situation is considered where previously defined relations $\mathbf{r}^1$ and $\mathbf{r}^2$ on $M \times M$ (where M is the entire domain), that give rise to two class-relations, each of one of the nine types $\exists\exists$, $\exists\forall$, $\exists\forall_o$, $\forall\exists$, $\forall\forall$, $\forall\forall_o$, $\forall_o\exists$, $\forall_o\forall$ and $\forall_o\forall_o$ as defined in table 3.2. I.e. we have that $A\, r^1_{HI}\, B$ and $B\, r^2_{JK}\, C$, where H, I, J and K are quantifiers with $H, I, J, K \in \{\forall, \exists, \forall_o\}$.

The composition of these class relations, denoted by $\odot_B$, is then:

$$(A\, r^1_{HI} B) \odot_B (B\, r^2_{JK} C) \equiv A \left(r^1_{HI} \odot_B\ r^2_{JK}\right) C. \tag{3.11}$$

Our task at hand is now to determine which type, if any, this class-relation $r^1_{HI} \odot_B\ r^2_{JK}$ between A and C is: Given (3.11), we will investigate whether there exist quantifiers $L, M \in \{\forall, \exists, \forall_o\}$, such that $A\, r^3_{LM}\, C$, where r^3 is the "lifted" relation of $\mathbf{r}^3 = \mathbf{r}^1 \circ_B \mathbf{r}^2$. $\circ_B$ is here a modified version of the conventional composition of relations, where we restrict the mediating element m to be in a specified, non-arbitrary subset of M:

$$\mathbf{r}^1 \circ_B \mathbf{r}^2 = \\ \{(m_1, m_2) \in M \times M \mid \exists m \in B(m_1 \mathbf{r}^1 m \wedge m \mathbf{r}^2 m_2)\}.$$

This B-composition corresponds to the intuition behind the knowledge representation using classes which are based on complex role inclusions using individuals.

An example of such B-compositions is:

$$A\, regulates_{\forall\forall}\, B\ \odot_B\ B\, isa_{\forall\exists}\, D \sqsubseteq A\, regulates_{\forall\exists}\, D,$$

whereas a non-B-composition as $A\, r^1_{\forall\forall}\, B\ \odot\ C\, r^2_{\forall\exists}\, D \sqsubseteq A\, r^3_{\forall\exists}\, D$ often is an un-acquired composition for knowledge modeling purposes.

To illustrate our task, the example given in section 3.2.1 is considered, i.e. we consider given relations $A\, r^1_{\forall\exists} B$ and $B\, r^2_{\forall\forall} C$. The next step is to find

Table 3.3: The resulting relationship types of the composite of two class relationships $A\ r^1_{\exists A}\ B\ \odot\ B\ r^2_{\forall A}\ C$ [4].

Type of r^1	Type of r^2										
	°A°A	A°A	E°A	°AA	°AE	AA	AE	EA	EE		
°A°A	°A°AℨEE	A°A	A°A	°A°AℨEE	A°A	A°AℨAE	A°A	A°AℨEE	A°A		
A°A	°A°AℨEE	A°A	A°A	°A°AℨEE	A°A	A°AℨAE	A°A	A°AℨEE	A°A		
E°A	°AA	—	—	°AA	—	°AA/AE	—	°AA/EE	—		
°AA	°AA	—	—	°AA	—	°AA/AE	—	°AA/EE	—		
°AE	°AA	—	—	°AA	—	°AA/AE	—	°AA/EE	—		
AA	°AAℨEA	A°A/EA	A°A/EA	°AAℨEA	A°A/AE	AA	AA	EA	EA		
AE	°AAℨEE	A°A/EE	A°A/EE	°AAℨEE	A°A/EE	AE	AE	EE	EE		
EA	°AA	—	—	°AA	—	AA	—	EA	—		
EE	°AA	—	—	°AA	—	AE	—	EE	—		

out whether $\mathbf{r}^3$ defines a class-relationship between classes A and C, where

$$\mathbf{r}^3 = \Big\{(x,z) \mid \Big(\exists y\big(B(y) \wedge x\mathbf{r}^1 y \wedge y\mathbf{r}^2 z\big)\Big) \wedge \Big(\forall x\big(A(x) \rightarrow \exists y\big(x\mathbf{r}^1 y \wedge B(y)\big)\big)\Big) \wedge \Big(\forall y\big(B(y) \rightarrow \forall z\big(C(z) \rightarrow y\mathbf{r}^2 z\big)\big)\Big)\Big\}$$

If so, the given class-relationship will then be $A\ (r^1_{\forall\exists}\ \odot_B\ r^2_{\forall\forall})\ C = r^3_{LM}$. According to table 3.3, in this case we have that $L = M = \forall$, i.e. $A\ (r^1_{\forall\exists}\ \odot_B\ r^2_{\forall\forall})_{\forall\forall}\ C$.

Further proofs concerning $\mathcal{CRL}$-compositions can be found in [4].

3.3 Summary

In this chapter, logical, knowledge-base formalisms such as $\mathcal{FOL}$ and subsets of this were introduced, including newly developed relation-based logics known as class-relationship logic and description logic.

Additionally, a revision of [4] focusing on the contribution of this author is presented, suggesting possible compositions of class-relationships, which are shown in table 3.3.

Chapter 4

Knowledge representation II: Domain knowledge

In every knowledge engineering task, knowledge of the domain of interest is crucial. Since this researcher's role is both that of knowledge engineer and biochemistry consultant, an analysis of the domain knowledge is a natural starting point for delving further into the formal properties of biomedical regulatory relations.

Extraction of biomedical relations and biomedical events in texts has been investigated broadly in the last decade, due to the role of regulation in both biochemical pathways, drug development and other areas within molecular biology.

The approach in this chapter is to investigate the area of regulation based on the intensional semantics of the concept, however, not so much on the many different regulations such as "translational modifications," "protein-protein interaction" or "MAP-kinase pathways." The relationships of both negative and positive regulates are explained qualitatively as well as formally, using representation forms from kinetics, graph representation and logical semantic formalization.

First, the biological knowledge that creates the platform, on which reasoning rules, semantic models and lexical frames will be built, will be introduced. Sections 4.1-4.3 contain an extended revision of the domain description in [136], in which the domain for the purpose of describing a formal semantics of regulation as relation was analyzed.

Next, section 4.4 treats a further formal semantic analysis, also based on [136]. This is based on the biological knowledge from the beginning of the chapter as well as the formal foundations of chapter 3.

Lastly, the development of formal reasoning rules of regulates as relation, and on composites of these, is presented in section 4.5, which is based on [134, 135] and includes work from chapter 3.

4.1 Biology of regulation

Molecular regulation is used on a large-scale within the field of systems biology. In this area, regulatory genomics, and regulatory networks in general, contain knowledge about different substances such as small molecules and gene products and how they interact with each other. For the purposes of this dissertation, the following broad statement is suggested:

> *A regulatory event is characterized by one or more molecules regulating one or more processes or molecules. The molecules can be gene products (proteins or functional RNA), molecules generated by some process or introduced to the organism by food intake or similar means. These processes can then either physically or chemically alter other molecules (e.g. DNA strands, promoters, enzymes, etc.) and, thereby, activate or block other processes.*

Knowledge extraction of these events can lead to new theories on dynamic regulatory mechanisms within an organism or a family of species such as vertebrates.

Other than the field of biochemistry, the regulates relations: *regulates*, *positively regulates,* and *negatively regulates*, are central in for example in economics. In economics, "a growth in industrial production will stimulate the rate of inflation" and "growing real interest rates will cause a down regulation of the rate of inflation" [81] can be approximated in a simple model concerning macro-economics.

In biomedical pathways, which are the focus of this dissertation, it is typically gene products and smaller molecules that interact with each other in complex processes using approximations like the economics examples above.

This could be demonstrated in a *knowledge pattern* as described by [34], where knowledge on regulation in biology, e.g. the work in section 4.5, could partly be reused for modeling knowledge of economics or other dynamic systems. A part of a *knowledge pattern* is a domain pattern, which reflects knowledge within a domain that can operate on different sources belonging to the domain, e.g. biomedical regulation, and perhaps later may be adapted to other domains.

4.1.1 Examples of regulatory events

A commonly used example of regulation is the insulin response mechanism. Insulin stimulates, through a long regulatory path, uptake of glucose through cell walls, protein synthesis and glucogenesis, among other functions. These stimulations occur via e.g. an activation of the gene product PIP3, and the glucogenesis is triggered through an activation of the Akt protein and inhibition of PHK as shown in figure 4.1. Note here, that most of the regulatory paths are between biological entities, gene products and small molecules,

typically meaning one gene product regulates the level (by the production or secretion) of another gene product.

Another example of how knowledge of a bio-pathway has been investigated semi-formally for hypothesis testing is the presentation of the damage response pathway in yeast [132]. This thesis presents an example of applications in a laboratory strategy for predicting a longer "pathway traveling" with respect to gene products which regulate the production of other gene products. The methodology is not formally defined, but it has an algorithmic nature:

In [132], the notion of "deletion-buffering" is used. The meaning of this is: when a transcription factor X, of the gene product G is removed, it results in the inability of the gene product to regulate the production of another protein P. This typically results in an activation of P if G is an inhibitor and an inhibition of P, if G is an activator, and G interacts directly with P. Thus, by "deleting" a transcription factor for G, a higher production of P is obtained.

Compared to this study, the semantic analysis of section 4.4 generalizes the understanding of regulates. This work is not focused on whether the agent of the regulatory event is a transcription factor. As long as the factor has a positive or negative regulatory effect on another factor, they are basically in the same class/of the same kind (with same general type of individuals). In section 5.4, lexical text patterns will be presented to explore the more general semantic patterns in section 4.4.

There will of course be examples that are non-trivial compared to the insulin-regulation example. An example is the recently discovered miRNAs, which are small regulating transcripts. miRNAs typically regulate gene products by binding to the mRNA of the gene product. However, often the regulation is just predicted *in silico* by sequence analysis until the experimental data has verified (or falsified) the interaction. This information is difficult to model by a simple regulates-relation since the meaning is rather *predicted_ to_ inhibit*. This vagueness, as well as other uncertainties, are discussed further in section 7.3.1.

4.2 Representation of regulation

Knowledge on biomedical pathways is typically represented in the expressiveness span from simple graph representations among gene products to the more sophisticated linked differential equations which take detailed information like rate and level into account.

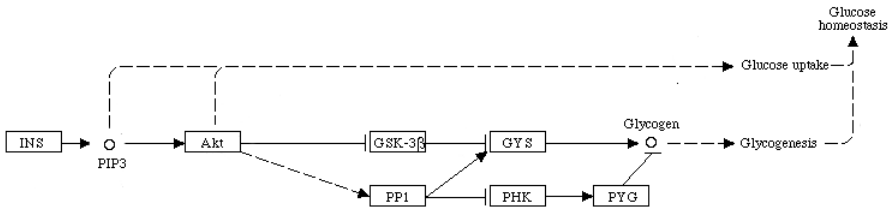
4.2.1 Regulatory webs: Graph representations

Regulatory networks are represented in databases like KEGG and Reactome in relatively simple structures, containing simple information on whether a substance up or down regulates another substance or process [91, 73].

Interactomes, created by laboratory experiment resources like [74, 65, 40, 78], are even simpler containing only the information that something interacts with something else. This interaction information can either be a regulation between two substances or a collection of smaller proteins creating a protein complex and Interactomes can accompany in knowledge acquisition of biomedical pathways.

Graph representations are mostly informal and made to illustrate a regulatory path. In more formal graphs like KEGG [73] (in figure 4.1), Reactome [91], and metacyc [31], regulation among entities like small molecules and gene products are formalized into a database, into which simple queries may be made. For example, in the network of figure 4.1, it is displayed that "PP1 activates GYS" and by the legend it can be inferred that "PP1" is a gene product, which is information stored in the relatively simple structure of KEGG. This figure will be returned to in later sections, since it serves as an illustrative example of a fragment of a regulatory event.

Figure 4.1: In KEGG, a regulatory relation is represented by either an arrow which refers to up-regulation, or by an arrow with an orthogonal line corresponding to down regulation. *Akt* activates *PP1*, which inhibits *PHK*, which activates *PYG*, which inhibits the *Glycogenesis* process. Overall *Akt* activates *Glycogenesis* through three different paths. The figure is reduced compared to the original one (ID:04910, Appendix A) [73].



4.2.2 Linked differential equations

At the other end of the scale, regulatory pathways can be represented using linked differential equations on the form

$$\frac{dS_i}{dt} = f_i(S_j, p_k) = f_i^+ + f_i^-, \quad \begin{matrix} i,j = (1,...,n) \\ k = (1,...,m), \end{matrix} \tag{4.1}$$

where S_i is the biological entity influenced(e.g. gene product or small molecule), t is the time, p a kinetic parameter proportional to the concentration of the causal agent, n is the number of biological entities influenced and m is the number of the kinetic parameters. $f_i^+$is the sum of the incoming flux (leading to positive regulation if larger than f_i^-) and f_i^- is the sum of the outgoing flux leading to negative regulation if larger than f_i^+ [64].

This representation is very expressive and difficult, if not impossible to implement with the almost 10.000 regulations [31] represented in KEGG.

Linked differential equations are widely used within physical chemistry and biophysics to simulate dynamic regulatory changes within a fragment of a regulatory network. Storage of kinetic parameters for a certain route of a pathway together with the equation (4.1), can be referred to as a knowledge base using the well-known, and potentially expressive, functional algebra [43] rather than first-order predicate calculus.

Example of Michaelis-Menten. A simple example of a regulatory mechanism can be found in Michaelis-Menten kinetics in which an example on an enzymatic function:

$$A + B \underset{k_{-1}}{\overset{k1}{\rightleftarrows}} AB \rightarrow A + C. \tag{4.2}$$

Here, A is the reacting enzyme that binds to B and, thus, moderates B into C. The way this regulatory mechanism can be described is complex. Either the interpretation is that B is activated by becoming C, or B is inhibited by removal or breakdown.

A simpler example is the reaction:

$$A + B \underset{k_{-1}}{\overset{k1}{\rightleftarrows}} AB. \tag{4.3}$$

From this, it can generally be inferred that A inhibits B (since it binds to B into a new complex) and vice versa.[1] To examine the rate and estimate whether it is an inhibition or an activation, the calculation is:

$$\frac{d[AB]}{dt} = k_1[A][B] - k_{-1}[AB], \tag{4.4}$$

[1] In the cases where the complex AB can be seen as a blockage of B.

i.e. how the product AB is influenced by the enzyme A. The Michaelis constant is in the simple case $K_M \sim \frac{k_1}{k_{-1}}$ and the larger a positive value of K_M, the higher the stimulation of the product $[AB]$ and the faster the pacification of B until equilibrium is obtained.

4.2.3 Logical representation

Within the field of computer science and systems biology, an approach to the representation of biological regulatory networks has been investigated in the last decade [63, 43, 136]. In this field, the approach is qualitative, but still more expressive, than the simple graphs described earlier in this section.

In [43], the authors claimed that abstract models are important for discrete reasoning. A model and allowable reasoning steps are formally defined and predictive power in hypothesis support provided. In this ontology, amino acids like serine, tyrosine and threonine are constants of the sort *Amino acid*, just as EGFR is a constant of the sort *Protein*.

Compared to the differential equations described earlier in this section with [63], the following definitions are suggested. Given a set of object parameters (x, y, z) to be functions of time and K_M to be the Michaelis constant as defined above, the qualitative constraints corresponding to regulation are:

$\left\{\frac{d}{dt}x(t) < 0\right\} \sim \{K_M < 1\}$ (negatively regulated x),

$\left\{\frac{d}{dt}x(t) > 0\right\} \sim \{K_M > 1\}$ (positively regulated x),

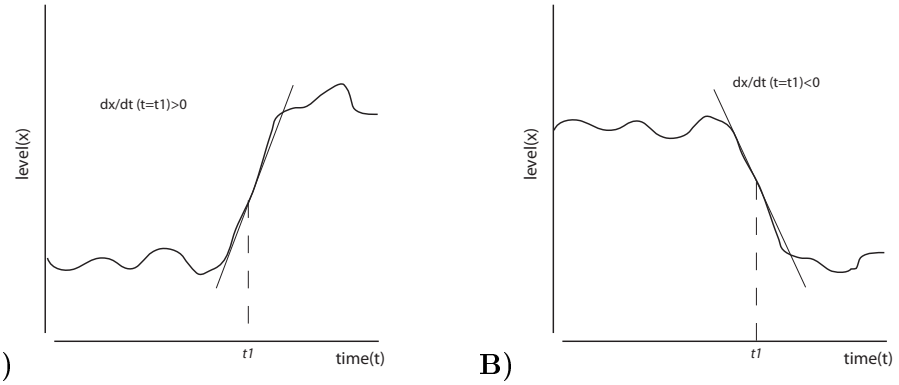
$\left\{\frac{d}{dt}x(t) = 0\right\} \sim \{K_M = 1\}$ (constant or steady state of x - no regulation).

The main difference between representation in linked differential equations and in logics is the means of moving from describing a system with continuous variables to a discrete description, such as that of e.g. Biosim [63]. In Biosim, several parameters are used, although the only values (qualitative magnitudes) are $\pm\{0, 1, inf, std\}$.

However, this thesis will illustrate the connections between logical quantifications and rules and the differential equations at the base of the understanding of a regulatory process. In figure 4.2, a basic understanding of inhibition and activation is illustrated.

Notice that the overall gradient counts, since it is common to have local fluctuations within a "stable" process. This can be seen from the figure 4.2, in which a simple case of up and down regulation is illustrated.

Figure 4.2: Inhibition and activation exemplified by a graph for the level x as a function of time, t ($x(t)$). A) shows a process that is stimulated, having a gradient larger than zero $\frac{d}{dt}x(t) > 0$ at $t1$, and B) shows an inhibition ($\frac{d}{dt}x(t) < 0$ at $t1$).



4.3 Ontological assumptions about regulation

Based on the examples in section 4.1.1, the following will investigate the ontological aspects of regulatory relations as they present themselves in biomedical research in more detail. The entities serving as subjects to relationships will be clarified and a distinction between general concepts or classes and individuals instantiating these classes will be made.

4.3.1 Research practice and granularity

When creating models of knowledge to be used in biomedical hypothesis development, inspiration can be obtained from the practice in the biomedical laboratory. Before the validation of a hypothesis is carried out in the laboratory, the precise rates and levels of the substances involved are not necessarily the first aspect to consider. Rather, a qualitative overview of the processes is the primary instrument at this stage of discovery.[2] The precise levels of substances are generalized or ignored and whether levels are affected positively or negatively is the key concern, leaving a higher level of abstraction (as in [43]).

Likewise, what is really meant by the statement, "insulin positively regulates glucose transport", is that an amount (or pool) of insulin causes an amount of glucose molecules being transported. Or, expressed more precisely

[2] In later stages, considerations like level and other laboratory experiment settings are taken.

with respect to laboratory experiments: When the level of insulin rises, this rise causes a higher frequency of glucose molecules to be transported (through cell walls). This rise is typically directly or indirectly caused by the external addition of a substance - either by intake of nutrition, medication or even by lack of intake of the necessary nutrition to make the system work properly. Hence "amount," "pool" or "aggregate" of substances are the basic entities, or "individuals," subject to possible relations.

These individual types count for substances. Introducing organs, or larger functional entities of the body, call for a another granularity than "amounts." Hence, an entity is rather one heart or one pancreas, etc.

Similar concerns about the difference between aggregates and individual objects have been discussed in a work within the top-level ontology BioTop [117]. In this work, granularity was investigated for the categories in a biomedical context. Within this frame, exactly the issue of Object vs. ObjectAggregate as classes in Basic Formal Ontology (BFO) was investigated:

> *"BFO's alleged assumption that there is a clear ontological division between Object or ObjectAggregate is already challenged by the fact that any self-connected physical object can be also be described as a mereological sum of molecules, atoms, or elementary particles." [117]*

What the researchers stressed is that if *insulin* is seen as an Object instance, it cannot be an ObjectAggregate instance at the same time. A solution suggested is to create classes like PortionOfGas or PopulationOfHumans and to also have a top-level, class-like Population and Portion (or Amount), as "children of the union class" [Object OR ObjectAggregate] [117].

The approach of this work is a bit different; instead, it is suggested that any individual within a regulatory network be conceived as an amount, eliminating the need for extra classes. This is also possible to implement by using ontological constraints of the relata of a regulates relation, such that only selected top-level types are admissible for containing "amounts" or "aggregates" as individuals.

4.3.2 Underlying assumptions on instances and classes

The former subsection suggests that amounts of molecules rather than single molecules are the central concern of biomedical researchers. In a laboratory context, researchers operate with certain amounts or batches of fluids containing multiple molecules. Also, the organs of the human body secrete an amount of molecules for regulating processes. Nothing really occurs if *one* beta cell secretes only *one* insulin molecule.

Thus, although it seems awkward to call an individual or basic entity, e.g. an "amount of insulin molecules", an amount will have very different properties compared to one single insulin molecule, which, in a nano-lab

setting might make sense; but not in the body nor in interaction experiments, where at least thousands of molecules are present.

The ontological assumptions made here are virtually in line with that presented in [119] and [121]. As in [119], classes and instances will be distinguished.

Here, classes refer to what generally exists, such as insulin, glucose, glucose transport, stem cell, etc. In the following; names of classes will be italicized and begin with a capital letter, for instance $Insulin$ and $Glucose\_transport$.

The distinction between classes and instances allows analysis of a natural language expression, such as "insulin positively regulates glucose transport," in more detail. A certain relation between individuals of the classes $Insulin$ and $Glucose\_transport$ exists, as presented in section 4.3.1.

Thus, relations between the instances are assumed to be a given, for example, by experimental evidence in the laboratory. On the basis of these, relations among classes or concepts are defined. For instance, "positive regulation" relations may exist among particular amounts of insulin and particular glucose transports. On the basis of these, a relation between the classes $Insulin$ and $Glucose\_transport$ can be identified. This will occur in section 4.4.

The notation in this chapter is the same as that in chapter 3 and an overview of the notation is given in table 3.1.

4.4 Analysis of logical semantics of regulation

In this section an initial semantic framework on an abstraction of the biological notion of regulatory pathways using logics is applied based on work in [136]. This will link some of the formalisms presented in chapter 3 and the domain knowledge and ontological assumptions presented in this chapter.

The focus is on the relations of *positively regulates* and *negatively regulates* as well as *regulates*, which is assumed to be a super relation of the two others (argumentation of this is found in chapter 5). The terms *stimulates* and *inhibits* are used interchangeably with *positively regulates* and *negatively regulates*.

The aim of a logical knowledge representation in this analytic step is to capture the formal semantics of the relations. Furthermore, logic implementations offer an opportunity to reason automatically (in a qualitative way) with the goal of obtaining new knowledge. This representation can be utilized in further work on lexical-semantic annotation to be used in information retrieval systems for example. Additionally, the representation can simulate regulatory networks in biology in a relatively simple manner.

Logic has always played an essential role in knowledge representation. In relational databases, a logic lies underneath and natural language contents can be expressed (to some extension) in logic [83], which has been investigated over the last couple of hundred years.

As a newer example, the popular language for semantic web, OWL, has a semantic based on description logics, described in section 3.1.2. Another classical logic for knowledge representation is first-order logic (described in section 3.1.1), a fragment of which is description logic. First-order logic is used since it is more expressive, but also the possibility of using description logic will be discussed.

Notation. In line with [119], continuants (or more precise, substances, corresponding to gene products, small organic molecules and ions) and processes referring to the top-ontology BFO [122] are distinguished.[3] Continuants are entities that continue to exists over time and may undergo changes, contrary to processes, which are events. Continuants are entities that can change and the changes themselves are processes. The representation $c, c_1, c_2, ...$ will be used to range over continuants and $p, p_1, p_2, ...$ to range over processes. An example of a continuant in the domain of this thesis could be an amount of insulin, whereas a glucose transport is a process.

4.4.1 A logic formalization of regulatory relations

Given the ontological assumptions in section 4.3, the possible relations between classes involved in the knowledge represented are discussed. First-

[3] In the linguistic analysis of section 5.4.4 ontotypes from UMLS was used instead.

order logic is used in the definition of relations among these classes.

An example of a formalized relation between classes is the "part of" relation present in many biological ontologies. One can state that *Cell membrane* is "part of" *Cell*, which expresses the fact that every particular cell membrane is the membrane of a particular cell. In other words, "for every cell membrane there exists a cell of which it is part of". Assuming a **part_of** relation between individuals, one can define a $part\_of$ relations among classes C_1 and C_2 in the following way [121]:

$$C_1\ part\_of\ C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \exists y(x\ \mathbf{part\_of}\ y \wedge C_2(y))). \tag{4.5}$$

A relation between classes defined this way will be called "$rel_{\forall\exists}$" in line with the notation in table 3.2. Generally we define a $\forall\exists$ relation $rel_{\forall\exists}$ between two classes, C_1 and C_2, based on a relation **rel** between individuals, by:

$$C_1\ rel_{\forall\exists}\ C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \exists y(x\ \mathbf{rel}\ y \wedge C_2(y))). \tag{4.6}$$

The two classes C_1 and C_2 are also called the *relata* of the relation. Another example of a concrete relation between classes is the one exemplified by the term "enzymes stimulate processes". Even though it may not be visible on the surface, an even stronger tie between the two classes is present than what is expressed by a $rel_{\forall\exists}$ relation. The relationship between enzymes and process is such that whatever an enzyme stimulates, it is a process. A relation of this kind will be called "$rel_{\forall\forall o}$" (inspired by the Manchester syntax). Formally we define the $rel_{\forall\forall o}$ relation by:

$$C_1\ rel_{\forall only}\ C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \forall y(x\ \mathbf{rel}\ y \rightarrow C_2(y))). \tag{4.7}$$

Consider the case $positively\_regulates$ as exemplified by a phrase such as "insulin positively regulates glucose transport", which exemplifies the kind of knowledge we aim to represent. As previously discussed in section 4.3.1, a sentence like this should be read as "for all amounts of insulin and all glucose transports, the insulin can potentially positively regulate the glucose transport". To express a relation like this we introduce the "$rel_{\forall\forall}$" relation between classes in the following way:

$$C_1\ rel_{\forall\forall}\ C_2 \quad \text{iff} \quad \forall x(C_1(x) \rightarrow \forall y(C_2(y) \rightarrow x\ \mathbf{rel}\ y)). \tag{4.8}$$

The reason for choosing the $rel_{\forall\forall}$ relation instead of the $rel_{\forall\forall o}$ to represent knowledge such as "insulin positively regulates glucose transport", is that insulin also has the potential to stimulate other processes such as glycogen production. This potential is excluded if the knowledge is represented as a $rel_{\forall\forall o}$ relation.

4.4.2 Ontological types of relata

A deeper ontological analysis of the entities involved in regulatory relations reveals that a distinction between the top-level types of relata like continuants and processes has ontological significance. In this work, we have used classes from the top-level ontology BFO [122] which have also been used to define the biomedical relations of the Role Ontology described in [119].

Relations between individuals have to be divided into cases depending on whether the individuals are continuants or processes. However, in this section it is shown that this division can be collapsed using operators for production and output.

These triples $\{relata_1 - R - relata_2\}$ are similar to a lexical semantic representation as regulatory knowledge patterns described in [138] and in section 5.4, where the transformation assists in the semantic extraction of biomedical texts. Additionally, they have been described as Ontological affinities in Concept algebra as presented in section 3.1.4.

In case of regulation, continuants can regulate other continuants or processes, but processes can also regulate other processes or continuants. Thus, there are four possible regulatory relations depending on whether the related individuals are continuants or processes. Focusing on the relation "stimulates" there are the four relations: $\mathbf{stimulates_{cc}}$, $\mathbf{stimulates_{cp}}$, $\mathbf{stimulates_{pc}}$, and $\mathbf{stimulates_{pp}}$ where, for instance, the subscript "*cc*" means that it is a relation that can only hold between two continuants. However, introducing a "production of" and an "output of" operator makes it possible to reduce these four relations to only one.

The $production\_of(...)$ operator works on a continuant c by transforming it to the process that is the production of c. Similarly the $output\_of(...)$ operator transforms a process p to the continuant that is the output of p.[4]

With these operators the instance relations $\mathbf{stimulates_{cc}}$, $\mathbf{stimulates_{pc}}$, and $\mathbf{stimulates_{pp}}$ can be reduced to the $\mathbf{stimulates_{cp}}$ relation. These reduction are given by:

$$\begin{array}{lll} c_1\ \mathbf{stimulates_{cc}}\ c_2 & \text{reduces to} & c_1\ \mathbf{stimulates_{cp}}\ production\_of(c_2) \\ p\ \mathbf{stimulates_{pc}}\ c & \text{reduces to} & output\_of(p)\ \mathbf{stimulates_{cp}}\ production\_of(c) \\ p_1\ \mathbf{stimulates_{pp}}\ p_2 & \text{reduces to} & output\_of(p_1)\ \mathbf{stimulates_{cp}}\ p_2 \end{array} \tag{4.9}$$

These reductions reflect how the relations are used as verbs in sentences in biological texts, for example: "Insulin $\mathbf{stimulates_{cc}}$ glycogen", "insulin $\mathbf{stimulates_{cp}}$ the glycogenesis", and "insulin $\mathbf{stimulates_{cp}}$ the production of glycogen (through the Glyconeogenesis)" where the process glycogenese is equal to "the production of glycogen". Likewiseit is possible to formulate the sentences: "beta cell secretion $\mathbf{stimulates_{pp}}$ glycogenese" that can be

[4] This is problematized in the discussion.

reduced to "output of beta cell secretion **stimulates**$_{\mathbf{cp}}$ production of glycogen", where the output of beta cell secretion is insulin.

A more detailed view on reasoning processes including information on the ontological types of relata are presented in chapter 6. The connection to textual representation will be discussed in section 5.4.

4.4.3 Concluding remarks on the logical analysis

Based on an analysis of the biomedical examples, and our declaration of the ontological assumption, we have suggested that the adequate formalizations of *positively* and *negatively regulates* in first-order logic are represented by the formula $\forall x(C_1(x) \rightarrow \forall y(C_2(y) \rightarrow x \ \mathbf{rel} \ y))$. A description of the relata, the First-order formulas, and examples of *regulates*, *positively_ regulates* and *negatively_ regulates* are displayed in table 4.1.

Table 4.1: Formal definitions of three basic regulatory relations expressed as class-level relations [136]. *Relation and relata* are described by the ontological types from BFO. *Definitions* displays the $\mathcal{FOL}$ formalizations, and *Examples* contributes with PubMed-abstracts examples.

A. Regulates

Relations and relata	C_1 $regulates_{\forall\forall}$ $production\_of(C_2)$; C_1 and C_2 are continuants.
Definitions	$\forall x(C_1(x) \rightarrow \forall y(production\_of(C_2(y)) \rightarrow x$ **regulates** $y))$.
Examples	...nitric oxide pathway regulates pulmonary vascular tone... ...non-histone chromosomal proteins may modify gene expression... ...CREB regulates cyclic AMP-dependent gene...

B. Positively Regulates

Relations and relata	C_1 $positively\_regulates_{\forall\forall}$ $production\_of(C_2)$; C_1 and C_2 are continuants.
Definitions	$\forall x(C_1(x) \rightarrow \forall y(production\_of(C_2(y)) \rightarrow$ x **positively_regulates** $y))$.
Examples	...IPA stimulates insulin release... ...Ca(2+) influx stimulates exocytosis of secretory granules... ...MMP-7 activates the epidermal growth factor...

C. Negatively Regulates

Relations and relata	C_1 $negatively\_regulates_{\forall\forall}$ $production\_of(C_2)$; C_1 and C_2 are continuants.
Definitions	$\forall x(C_1(x) \rightarrow \forall y(production\_of(C_2(y)) \rightarrow$ x **negatively_regulates** $y))$.
Examples	...GLP-1inhibits glucagon release... ...lithium inhibits the enzyme glycogen synthase kinase-3... ...RSBX negatively regulates an extension of the RSBV-RSBW pathway... ...insulin secretion from the β-cell to reduce IRI responses...

4.5 Regulates reasoning and knowledge retrieval

Reasoning and knowledge retrieval are traditionally concerned with formal representations and can be interlinked with information retrieval in applications.

However, in this research, reasoning within knowledge representation plays a role in itself and is described in this section focusing on two different works, namely domain-specific reasoning rules within the area of regulation and reasoning possibilities for class relationships, linking to section 3.1.3.

This section plays a foundational part as a potential basis for AI applications, since knowledge on regulation and discussions on domain-based reasoning rules (in contrary to pure logical inferences like inheritances through a taxonomy) are included. These reasoning rules require a domain analysis and, therefore, this section is titled knowledge retrieval.

4.5.1 Biochemical pathway logic

Biochemical pathways function in quite a different way than signaling systems, such as electrical signaling. Contrary to many signaling systems, gene products are generally constantly produced via transcription in the nucleus. However, many areas of the DNA are blocked and gene products and other chemicals in the cell also inhibit each other through complex regulatory pathways. The formal understanding of this is described in section 4.2.

This leads to domain-dependent formal properties for the relations, as will be described in section 4.5.3 rather than domain-independent properties such as transitivity for *isa*-relations. An inhibition of another inhibition leads to an activation. This is a biochemical research logic that is often simulated using coupled differential equations as described in section 4.2.2. The disadvantage of this representation is that it can be very heavy in complexity and will not automatically take ontological information into account.

In the research area of biomedical ontologies, the work with formal relations has recently reached a level at which larger projects invite participation [118]. It has been suggested that the logic implications of the relations should be analyzed thoroughly [121]. By using description logic-formalism, with an expressivity from of at least $\mathcal{EL}+$, and the reasoning tool, CEL, relations can be treated as modules with complex inclusions on forms like $R \circ S \sqsubseteq S$ [16]. In this format, e.g. the transitivity for the *isa*-relation is formulated as $isa \circ isa \sqsubseteq isa$, i.e. if insulin is a protein and if a protein is a molecule, then insulin is also a molecule.[5]

This study is concerned with the formal properties of the two relations, positive and negative regulation relations. The relations have been investigated in corpora and in relation to their logical implications. For ease of

[5] In an OWL-implementation this chaining would be obscure since the transitivity of isa is a part of the inference machinery of OWL taxonomies.

reading, they are termed *activates* and *inhibits*, which is equal to the legend in KEGG [73].

4.5.2 Usage of regulatory relations in Gene Ontology

Recently and not yet published, The Gene Ontology (GO) Consortium developed properties for regulatory relations [51]. These relations will be presented here as they are important to the properties suggested in this work.

The regulatory relations are described by GO as the following:

> *"Another common relationship in the Gene Ontology is that where one process directly affects the manifestation of another process or quality, i.e. the former* regulates *the latter. The target of the regulation may be another process—for example, regulation of a pathway or an enzymatic reaction—or it may be a quality, such as cell size or pH. Analogously to* part of*, this relation is used specifically to mean* necessarily regulates*: if both A and B are present, B* ***always*** regulates *A, but A may not always be* regulated by *B."*[6]

This can be interpreted as a $rel_{\forall\forall}$-relation, described in table 3.2:

$$\forall x(B(x) \rightarrow \forall y(A(y) \rightarrow x \,\mathbf{reg}\, y)), \tag{4.10}$$

given that A and B are present in the same place at the same time. This is equal to the findings in [136] which are discussed in section 7.2.2. However, due to the Gene Ontology formulation *"but A may not always be regulated by B"* [51] will be implicit in the $rel_{\forall\forall}$-relationship since it does not demonstrate a restrictional closure, such as the $rel_{\forall\forall_o}$-relationship.

Following the Gene Ontology Consortium, *regulates* have two sub-relations namely *positively regulates* and *negatively regulates,* which have similar formal properties. From this relationship, it can be inferred that if x *negatively regulates* y, then x also *regulates* y. The expression $isa \circ regulates \sqsubseteq regulates$ means: if A *isa* B and B *regulates* C, then A *regulates* C.

The reasoning rules of the relations of Gene Ontology are the following nine:

$$isa \circ regulates \sqsubseteq regulates \tag{4.11}$$

$$regulates \circ isa \sqsubseteq regulates \tag{4.12}$$

$$regulates \circ partof \sqsubseteq regulates \tag{4.13}$$

$$positively\_regulates \circ partof \sqsubseteq regulates \tag{4.14}$$

$$positively\_regulates \circ isa \sqsubseteq positively\_regulates \tag{4.15}$$

[6] geneontology.org/GO.ontology-ext.relations.shtml (2011)

$$negatively\_regulates \circ partof \sqsubseteq regulates \tag{4.16}$$

$$negatively\_regulates \circ isa \sqsubseteq negatively\_regulates \tag{4.17}$$

$$isa \circ negatively\_regulates \sqsubseteq negatively\_regulates \tag{4.18}$$

Note that in the Gene Ontology, which focuses mainly on event-concepts, no transitivity or anything similar is inferred through two or more regulatory relations [51]:

> *No inference is possible when a regulates relation is followed by a second regulates relation. This is also true for positively regulates and negatively regulates.*

As discussed in section 7.1.2, this is not necessarily against the reasoning rules provided in section 4.5.3. The Gene Ontology knowledge base is largely concered with processes, which are annotated to different genes, and not the direct regulatory reactions from gene product to gene product as in e.g. KEGG. Furthermore, a significant factor of uncertainty occurs within the reasoning steps, which might not be allowed in modeling the Gene Ontology, but which could be useful in heuristic pathway modeling.

4.5.3 Complex role inclusion of regulatory relations

Inhibit and *activate* are relations that, in a biochemical pathway demonstrate a special kind of inheritance, e.g. if x *inhibits* y and y *activates* z, then it can be deduced that x *inhibits* z, as formulated in $\mathcal{EL}^+$ using complex role inclusions [16]:

$$activates \circ activates \sqsubseteq activates \tag{4.19}$$

$$inhibits \circ inhibits \sqsubseteq activates \tag{4.20}$$

$$inhibit \circ activates \sqsubseteq inhibits \tag{4.21}$$

$$activates \circ inhibits \sqsubseteq inhibits. \tag{4.22}$$

The property of equation (4.21) is expressed in $\mathcal{FOL}$, for example:

$$\forall x \forall y \forall z \left(A(x) \wedge B(y) \wedge C(z) \wedge inhibits(x,y) \wedge activates(y,z) \rightarrow inhibits\,(x,z)\right), \tag{4.23}$$

where A, B and C may be different ontological classes. While *activation* is a transitive function, *inhibition* and *activation* are complex in combination and the binary property of the complementary pair can be formulated as a specific kind of relation.

The property of equation (4.20) is termed *inter-transitivity* (in multiplication, *two sided identity*). This property is written out in linked differential equations presented in section (4.2.2):

$$\left(\frac{d}{dt}x_3(t) > 0\right) = \left(\frac{d}{dt}x_1(t) < 0\right) \times \left(\frac{d}{dt}x_2(t) < 0\right), \tag{4.24}$$

corresponding to the simplified kinetic equation using Michaelis constants as described in section 4.2.2:

$$(K_{m3} > 1) = (K_{m1} < 1) \times (K_{m2} < 1), \tag{4.25}$$

which can be asserted about any relation having the properties shown in e.g. figure 4.2.

Similar in the combined property of equation (4.21) and (4.22), which can be called *cross-transitivity* (in multiplication it is termed, *left* or *right hand sided identity*),

$$\left(\frac{d}{dt}x_3(t) < 0\right) = \left(\frac{d}{dt}x_1(t) > 0\right) \times \left(\frac{d}{dt}x_2(t) < 0\right), \tag{4.26}$$

corresponding to the simplified kinetic equation,

$$(K_{m3} < 0) = (K_{m1} > 0) \times (K_{m2} < 0). \tag{4.27}$$

While the Gene Ontology definition of regulates as a relation does not allow these reasoning properties, [134] claims this can be of use for hypothesis testing and hypothesis development. The uncertainty within the steps of regulates relation should not be rejected, which will be discussed further in section 7.3.

4.6 Summary

In this chapter, the domain of regulation within molecular biology was discussed. An introduction to how biological regulation was represented in webs and how to understand the regulatory mechanism in differential equations was explained. Additionally, ontologies, including the conceptualization of regulation were presented.

In section 4.3 and 4.4, work from the paper [136] was presented, including a discussion of the ontological assumptions about regulates, the relata of regulates as a relation and a semantic analysis and formalization of the relation in $\mathcal{FOL}$ is given. In short, the relation was found to be described by the $\mathcal{FOL}$-formula: $\forall x(C_1(x) \rightarrow \forall y(C_2(y) \rightarrow x$ **regulates** $y))$ and the granularity of the relata x and y (if a *production_ of()* operator or similar operator is introduced) to be "amount."

Finally, work on reasoning within the domain is included in section 4.5. This is based upon the papers [135] and [134] and describes compositions of the regulates-relations formalized in a logical manner such as: *A inhibits B and B inhibits C -> A inhibits C.*

Chapter 5

Knowledge representation III: Linguistic analysis and modeling

A clarification of the fundamental view of ontology is not sufficient to create a useful ontology as the foundation of a knowledge base. The terminological and corpus linguistic aspect of explaining ontological concepts and relations are important for creating a coherent and consistent ontology, reflecting the tangible usage of the concepts in terms of the domain terminology.

In this chapter, modeling is examined, using basic principles and domain expert consulting for knowledge acquisition in section 5.1 based on [139, 38]. Next corpus analysis is treated in sections 5.2 - 5.4.

The corpus analytical section is separated into two:

- A statistical approach in section 5.3, based on [133, 134], primarily supports knowledge acquisition of domain specific verbs.
- A semantic and concordance analytical part in section 5.4, based on an extension of [138] as well as a fragment of [136], for the purpose of knowledge extraction and reasoning as described in chapters 6 and 7.

Roughly speaking, the corpus analytical approach is rather extensional in its resulting ontological description of the world, whereas the expert consulting and terminology modeling approach, on the other hand, results in an intensional ontology. In the last analytic section (5.4), a demonstration of a connection between intensionality and extensionality in ontology modeling is attempted.

5.1 Terminological principles

The terminology modeling of this section is highly focused on linguistic domain modeling and is based on methods presented in e.g. [89, 87, 86]. In short, we understand a terminological ontology as consistsing of concepts which are different from the top concept by a number of delimiting features. Practically, it is a modeling principle by which one delimiting factor is used for each subdivision in the specifications of the ontology. This procedure is specified below.

A terminological ontology is equal to a domain-specific ontology as used in the categorization of ontologies by Guarino [56], for example. In this section, we use the term terminological ontology as a synonym for the term, concept system, which is normally used in linguistic terminology work, e.g. ISO standard 704(2000) [1]. The method as described below is pointed out in algorithm 5.1.

In terminological ontologies, the main task is to reveal the terminology of a domain. This can be done by having nodes referred to as concepts which are described by means of concept relations. Characteristics that denote properties of individual referents belong to the extension of a concept.

All kinds of concept relations can be used used: type relations (ISA-relations), part-whole-relations and associative relations, such as causal relations. Characteristics of the concepts are presented as feature specifications in the form of attribute value pairs [29], e.g. INHIBITOR OF PROCESS: substrate (see figure 5.2).

On the basis of these feature specifications, subdivision criteria are introduced. The purpose of these is to provide an overview of the reasons for divisions and help the terminologist in writing consistent definitions. Subordinate concepts inherit the characteristics of superordinate concepts. A concept (with only one superordinate concept) may contain at most one delimiting feature specification. A concept (if not the top concept) must contain at least one delimiting feature specification. The notion of delimiting features is first found in Aristotle who mentions the differentiae [11] which is also used later by e.g. Peter the Great [123].

It is possible to introduce poly-hierarchy, i.e. one concept may be related to two (or more) superordinate concepts. The superordinate concepts should always belong to two different subdivision criteria. If this is not the case, the ontology should be changed.

According to the terminological principles, two concepts must not differ with respect to more than one characteristic, except if they belong to a poly-hierarchy, where the concepts in question have two or more superordinate concepts belonging to different subdivision criteria.

In our work, we use an iterative process folowing the algorithm 5.1: analyzing the concepts as well as placing them in draft concept systems in the form of hierarchies or networks on the basis of their characteristics. Then

Algorithm 5.1 Terminology modeling overview [89, 38, 139]. Below we describe the methods used to construct formal terminological ontologies, containing poly-hierarchy.

a. Find sibling concepts related to one superordinate concept.

b. Identify the characteristics of the concepts.

c. Can the sibling concepts be separated by one characteristic? If yes, introduce an attribute-value pair on each concept.

d. Group the siblings by means of one or more subdivision criteria.

e. If step c-d are not possible and there is a need for more delimiting characteristics on each concept, introduce an extra layer of concepts so that the sibling concepts form part of a poly-hierarchy, i.e. inherit characteristics from two (or more) superordinate concepts belonging to two (or more) different subdivision criteria.

f. Define the concepts as classes in e.g. OWL-DL. Create relations using "object properties" and subdivision criteria by "data properties".

g. Define the delimiting features of the sibling concepts by means of the logical equivalence operator. If a poly-hierarchy is present, the super classes are added as equivalents.

drafting definitions, and, finally, refining concept systems as well as definitions. In this respect, we progress to consistent definitions referring to the superordinate concept (i.e. genus proximum or nearest kind) followed by the delimiting characteristic. All concepts can thus be defined as the top concept + delimiting characteristics.

In the example shown in figure 5.1 (2), the genus proximum is *process*, the subdivision criterion or attribute is INFLUENCE, and one of the attribute values is "negative". The superordinate concept and the attributes of the feature specification must be the same in definitions of subordinate concepts falling under one subdivision criterion.

We suggest that the ontology/terminology modeling procedure is implemented as an iterative process. In the next section, we present examples of the steps that were used for constructing the Inhibition ontology as well as the micro ontology in the SIABO project. If this procedure is followed, the resulting ontology will have a minimum of necessary and sufficient conditions. It will consist of defined classes rather than primitives.

The procedure might be a template for later automatic conversions from terminology tools to Protege-OWL.

5.1.1 Examples of modeling

The outcomes of using the terminological principles within the work of this thesis are the following:

SIABO-domain ontology

A domain ontology, utilized by ontology-based search within the work of the SIABO project, also described in section 6.1.1. The ontology was modeled in collaboration with librarian Steen Christensen from Novo Nordic with an information retrieval scope. Information retrieval systems that utilize the ontology are described in [9] and [10].

The figure 5.1, displays both the modeled ontology and the basis of the generative ontology that querying and semantic indexing should be built upon.

Enzyme Inhibition in OWL

This ontology on (enzyme) inhibition as a concept is developed in collaboration with an enzyme chemist and a general chemist from the Danish Chemistry Society. The main scope of this ontology was to receive coherent descriptions of the concepts, as shown in figure 5.2.

Additionally, the ontology of figure 5.2 is implemented in OWL-DL using Protégé 3.4 ([53, 66]). The OWL file can be found at *ruc.dk/~sz/Inhibition09.owl*.

OWL-DL is used for its potential in modeling a fine grained property structure using e.g. the *hasValue* operator for data-type properties and the possibility of more functions in later extensions. For simplicity, two kinds of OWL-properties are used in order to represent concept relations, and feature specifications, as mentioned in section 3.1.4. Type relations and part-whole relations have an obvious formalization in OWL as *ISA* relations among classes and the so called object properties, respectively, which can be written like *isPartOf* in the recommended notation. In addition to these, we need to decide which type of property to use for the implementation of the feature specifications. In the present implementation, the features themselves are the data literals "strings of characters" that are inherited throughout the ontology. Therefore, we have chosen data-type properties to formalize the feature specifications to avoid introducing all the values of the feature specifications as classes.

As an example, see the string "Substrate" in SubstrateInhibition in figure 5.3: The class SubstrateInhibition has the value "Substrate" for the data-type property: *hasInhibitorOfProcess*. This property is inherited through the type relations and every class has exactly one value for each property. Any feature specification can be represented as a relation between two concepts, and a concept relation can be represented as a feature specification.

Figure 5.1: Resulting ontology-example from the SIABO project [9]. The first figure is an extract of the generative ontology and the second is the corresponding domain ontology.

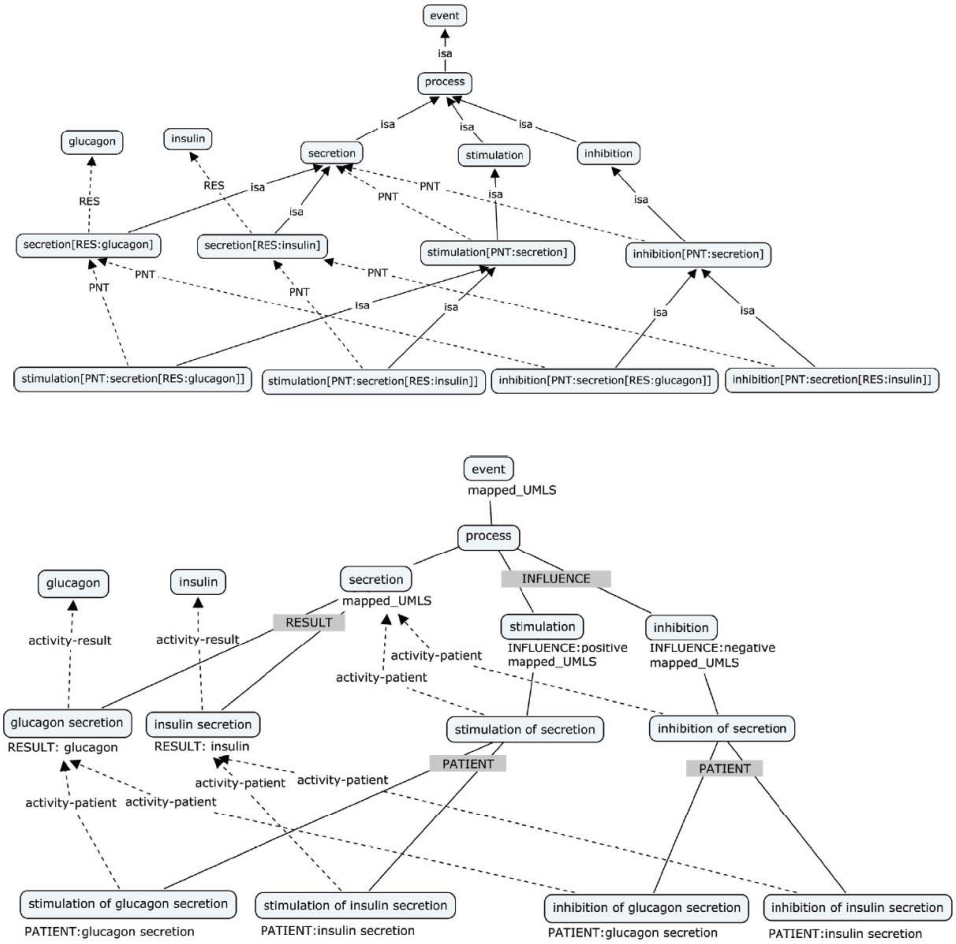


Figure 5.2: Domain ontology on inhibition as understood in enzyme chemistry [38].

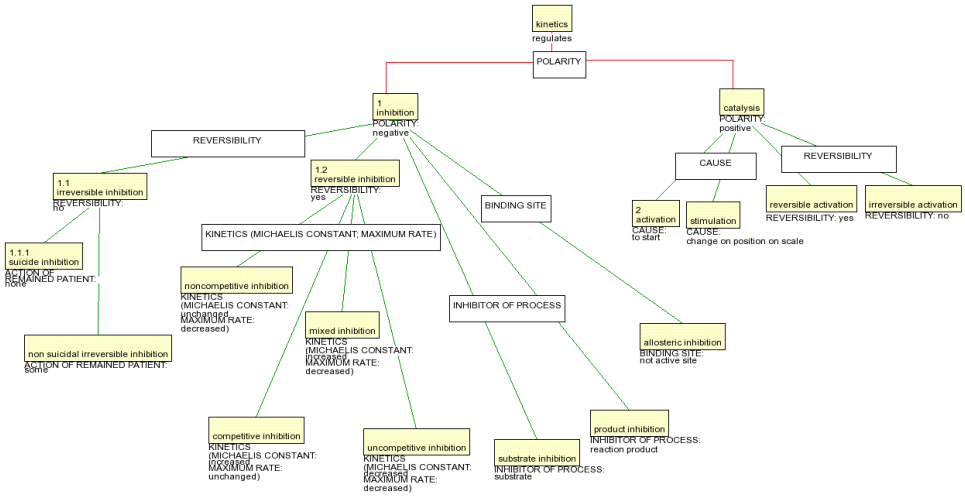
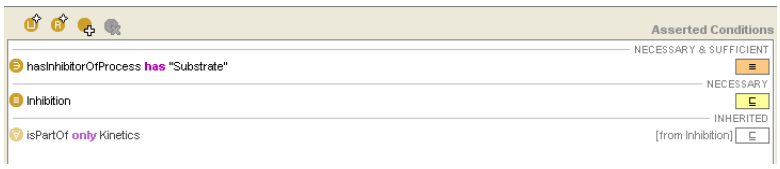


Figure 5.3: Conditions for the enzyme chemistry concept *Substrate inhibition* in OWL-DL [139].



Therefore, we could have considered using object properties instead, having the possibilities of creating transitive and symmetric relations. The full ontology does include such relations, namely *isPartOf* and *hasPart*, which can be transitive. Data properties are only inherited down in the hierarchy.

The principle of working with only one delimiting feature specification per concept becomes feasible in the formal modeling procedure: Siblings are all separated by characteristics, represented by feature specifications or "Data properties" in OWL. This supports a consistent ontology with a minimum of logical operators for each predicate since each concept can be described by its inherited characteristics and one "necessary and sufficient" description. (This is in line with the suggestion of Gruber's minimal ontological commitment [54].)

This modeling work, along with the OWL implementation is presented in [38] and in [139] in a modified $\mathcal{DL}$-form which is collected as formulars in Appendix D. It is displayed in a hierarchical tree form in figure 5.2 and the OWL document can be downloaded from *www.ruc.dk/~sz/Regrel/thesis*

5.2 Corpora and corpus analysis

The corpus analytic work of this thesis has been concerned with frequency lists for verbs, concordance analysis of verbs, and semantic text-patterns.

While the first (quantitative) discipline in section 5.3 supported ontology modeling, concordance (section 5.4) has been important in relation to the semantic analysis of the domain-specific relations and the latter as a framework for developing semantic frames that support an IR prototype.

Using domain corpora is an important aspect of the domain analysis for capturing how domain experts communicate and, thus, the extensions of their ontology. It helps to create a basis for domain modeling and understanding of the domain as a supplement to the intensional semantic analysis. Additionally, it provides a measure for later analysis of texts for the purpose of information extraction, semantic annotation and information retrieval as well as detection of semantic roles.

There are several corpora that can be useful for domain analysis within molecular biology. Most of them are based on Medline abstracts [93, 109], which are open access biomedical abstracts for academia and thus used by most of the NLP society within biomedicine.

In this thesis, the primary corpus sample is a Medline record consisting of 632.316 lines (around 40.000 abstracts). Additionally, a corpus consisting on 3.884 full biomedical patents from the diabetes domain and an extended corpus of 150.000 Medline abstracts, both selected by Novo Nordic for their domain relevance.

In addition to traditional NLP-methods as described in e.g. [105, 108, 6, 3], the corpus works in the rest of this chapter are inspired by the work on regulations based at the JULIE-lab at Jena University from 2008-2010 [27, 60, 28, 21]. Work from the JULIE-lab is introduced into this chapter (section 5.4) as well as in the related works in section 1.2, because the methodology, focus, and some results share similarity with this work.

5.2.1 Semantic roles and frames

Semantic roles are important for moving from syntactic text-patterns to semantic knowledge-patterns from texts. Fillmore called these roles "Case Frames," in which verb.argument relations holds across languages [45, 46]. Fillmore developed a set of six case roles (similar to the ones presented in the ontolog-framework of section 3.1.4), of which Agentive and Objective cases are equal to the meaning of Agent and Patient roles most frequently used in this thesis. In some works, the Patient role is also called "Theme".

5.3 Statistical corpus analysis

Within this work, a statistic corpus analysis has two main scopes: it helps identify concepts and relations for an ontology-modeling and it supports a comparison between the biomedical domain and general language, pointing out issues specific for the biomedical area.

In this section, frequency lists are used as an initial investigation of the usage of verbs representing the regulates relation.

5.3.1 Frequency lists

When choosing relations, examples that are specific for the biomedical area need to be identified, but need to remain not so specific that they cannot be used within other sub-areas of molecular biology than, for example, diabetes [133, 134]. Therefore, very specific verbs like methods for microbiological lab work (e.g. *immunoprecipitating*, *diluting*) were not considered.

Some other text mining approaches [32, 96, 65] focused on the pathway relations, the central relations that connect substances in biomedical texts; these are typically positive (a *activates* b) or negative (a *inhibits* b). The positive relation has the property of transitivity, whereas the negative one is more complex, though it still has a transitive-like behavior (as is described further in section 4.5.3).

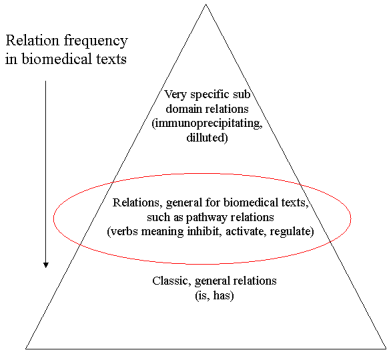
In addition to these domain specific verbs, the general, commonly used, and well studied relations represented by verb phrases such as *is a* and *has (the part)* will be used in the final ontology as well.

The aim, as a bioinformatician rather than a computer linguist, is to analyze the domain and point out the relations widely used within biology that can be grouped into a few fundamental semantic meanings. Figure 5.4 illustrates, roughly, the frequency of different kinds of terms in biomedical texts. Some are very specific for several methods (e.g. *dilute*, *immunoprecipitating*), some are general, well studied and highly frequent in many different texts (*is* in combination with *a*, *has* etc.) [37, 121].

Frequency of regulation in biomedical texts [133, 134] In between, are the frequent relations mentioned above that appears in most biomedical texts (*activate*, *inhibit,* etc.). These are highly represented in the frequency list of both patent verbs concerning biomedicine and a collection of biomedical texts when compared to the common language text. In addition to this researcher's experiments, other (search) tools concerning searches in biomedical literature have identified similar relation groups [32, 96, 65]. Thus, the initial focus was on the stimulatory and inhibitory relations, also proposed by others as important relations [32].

This study investigated verb frequency lists from Medline abstracts [93] and biomedical patents, comparing them with the general language corpus,

Figure 5.4: Overview of the frequency of verbs representing relations in biomedical texts. Some relations are general to all common language domains and some are general for biomedical texts [133].



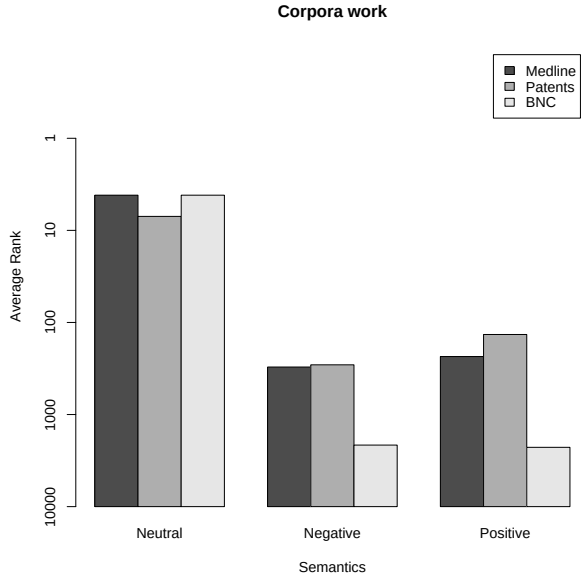
British National Corpus (BNC) [75]. The BNC is a standard reference corpus for the English language. It contains approximately 100,000,000 words, including 200,000 verbs. Medline abstracts are often used as a reference corpus in the biomedical area. However, it remains dynamic since new abstracts are added frequently. An arbitrary sub-section, with approximately 40,000 abstracts and 630,000 sentences was used. In addition to this, approximately 4,000 biomedical patents on diabetes and stem cells were used for the analysis.

Rough verb frequency lists, with un-lemmatized verbs using the first appearing form of the verb for all three corpora, were constructed. Of those, sets of verbs representing either negative regulatory relations or positive regulatory relations were manually chosen.

First, low-ranking verbs in the biomedical texts were manually inspected and, thus, detected that many low-ranked verbs (up to 500) had either the meaning of *inhibiting* or *activating*. The verbs were, and a search for the verbs in frequency lists executed. In addition, verbs with same semantics were looked up as used in Chilibot [32]. The final reduced set is shown in table 5.1.

The plots in figures 5.5 and 5.6 reflect the ranks of the verbs corresponding to the two different semantics in each corpora from table 5.1. The ten most common verbs from the BNC were also used as a background, which was called *neutral*. In figure 5.5 the average rank of the verbs in table 5.1 is displayed. The lower a rank is the higher the frequency. Figures 5.5 and 5.6 mirror the average frequency of the verbs on a logarithmic scale which can

Figure 5.5: The average rank of verbs with similar meaning from Medline abstracts, biomedical patents and the BNC corpus. *Neutral* means the ten most common verbs in the BNC, while *positive* means the verbs representing positive regulation and *negative* represents negative regulation [135, 134].



be used as an indication of differences between the corpora.

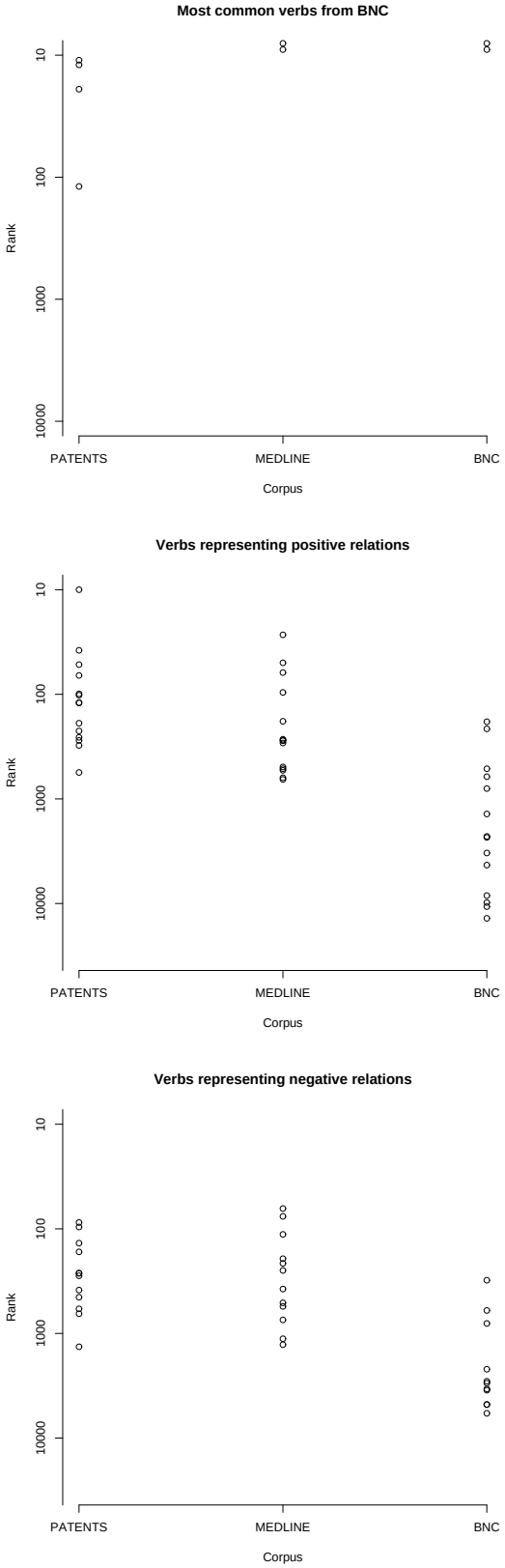
The background was constructed by the ten most frequent verbs in the BNC. Each had a similar rank in Medline, whereas the common verbs had a slightly lower rank in the patent texts. Verbs expressing both positive and negative regulation had an equally relative high frequency in the two biomedical corpora, where BNC had a much lower frequency of those verbs. The patents appears to have a weak preference for positive verbs compared with with those of the Medline abstracts.

The selected verbs and their frequencies can be found in table 5.1 [80] and figure 5.6. This indicates a couple of trends:

- It shows that most biomedical verbs have a much higher ranking (are much more frequent) in biomedical texts than the common language corpus (BNC), which could indicate that the verbs contains more information than other more common verbs and could be the subject for relations to be used by the developer when building the background ontology.

- Words that are common in normal texts (*start*) but have a specific meaning in biomedical context (*to activate* a process) have a much higher ranking in the BNC corpus than in the bio specific ones.

Figure 5.6: A plot of the ranks for each verb-meaning (as in figure 5.5) in the three corpora, Medline abstracts, biomedical patents and the BNC from table 5.1 using an inverse logarithmic scale for the y-axis [133].



- The type of domain text analyzed matters. The patents, though they only represent a narrow part of the biomedical area still have some differences in biomedical word frequencies compared to Medline and BioMed. For example, the words *encode, inactivate* and *remove* had a very low rank in the patents. A qualified guess for this issue is that many patent authors write texts in a cryptic general or non-informative way to make it more difficult to get information and thus make illegal copies, or competitors can gain important knowledge that is meant to be hidden, etc. The more typical biomedical verbs like *increase, induce* and *decrease*, are less frequent in the patent texts. (which the diabetes/stem-cell focus in the patent texts does not explain).

- The patents, Medline and BioMed are still considerably more similar than the BNC corpus, which is illustrated in figure 5.6.

Another interesting indication of the frequency lists is that the verbs identified by the Chilibot project [32] have a relatively high rank, which might mean that they are not used very frequently in either the patent texts or the BNC. However, these verbs have a bit lower ranking in the Medline and BioMed corpus (an open access part of Medline), indicating that they are somehow more frequent here, though not widely used.

Table 5.1: Verb ranks in biomedical patents, abstracts from BioMed Central, Medline and the BNC. R is short hand for "rank". If more forms are present the lowest rank is used [133].

Verb	Semantic relation	R (patents)	R (Medline)	R (BioMed)	R (BNC)
reduce	negative	87	76	47	310
remove	negative	96	250	487	603
inhibit	negative	137	113	297	4789
decrease	negative	166	64	91	4778
delete	negative	265	1281	1260	5806
regulate	negative	266	214	122	3391
block	negative	281	376	842	2987
limit	negative	385	193	161	803
inactivate	negative	451	1124	1293	NA
suppress	negative	582	510	900	3490
eliminate	negative	648	742	923	2198
attenuate	negative	915	736	796	17368
abolish	negative	1342	551	1325	2868
encode	positive	10	278	430	10704
express	positive	38	62	34	613
produce	positive	52	96	173	214
increase	positive	66	27	16	515
generate	positive	99	181	154	1394
secrete	positive	102	628	485	13853
induce	positive	120	50	83	3290
activate	positive	121	269	186	4291
amplify	positive	189	651	396	9795
stimulate	positive	224	292	426	2333
start	positive	276	515	482	183
promote	positive	308	495	357	796
facilitate	positive	258	533	229	2285
elevate	positive	559	275	331	8400

5.4 Concordances and lexico-semantic patterns

Only 30 percent of all co-occurring protein pairs in PubMed abstracts interact [5]. This fact motivates a fine-tuned semantic mining, on top of a simple co-occurrence analysis, for the task of finding interactions among proteins, other substances and processes.

For generating appropriate text patterns from a given corpora, a concordance analysis can be of help. A concordance is a list of occurrences of a certain word/text-item in a corpus of interest with an adequate amount of its surrounding context. The purpose of a concordance analysis often enables studies on how a word is used in context with respect to syntax, semantics and phrase structure on both sides of the word.

Key Words In Context(KWIC) is commonly used for a concordance, where the text item is aligned centrally and a certain amount of words surround the item (here, *regulates*) on the form:

left context	**key word**	**right context**
(...) , which possibly	*regulates*	the regional vascular tone.
(...) a nuclear substance that	*regulates*	the rates of estrogen dissociation.
evidence and theoretical arguments presented that phytochrome	*regulates*	the synthesis of new enzyme molecules against(...)

The list of concordances can be sorted with respect to right or left contexts, to provide a better overview of the textual forms [105] and provides the possibility of creating semantic text patterns. A glossary for these different text patterns is included in table 5.2.

Three outcomes from the syntactical and semantic analysis of the concordance emerged:

1. Generalized syntactical text patterns using semantic roles in the form of: [`Agent`] V-active [`Patient` Action-NN] used for machine learning in combination with 2) (section 5.4.3). These are called *patterns* or *(surface) text patterns*.

2. A shallow, semantic type annotation for each regulatory event identified from, e.g. the phrase: *"ethanol inhibits 3h-gaba release."* The relata, surrounding ontological types, are in this case: AGENT=ethanol=substance and PATIENT=3h-gaba release=process and the relation is within the frame "Negative regulation/Hindering" (section 5.4.4 and table 5.4). Constraints connected with the surface text patterns are termed lexico-semantic patterns in this thesis.

3. Partial semantic annotation of corpus text in the form of: AGT:[ethanol\NN] inhibits\VBZ PTN:[3h-gaba\NN release\VB] (AGT=substance,

PTN=Process) as a golden standard for comparing automatically generated semantic annotation.

A combination of 1 and 2 above is termed *knowledge patterns* or *frame parts*, whereas 3 is a *semantic annotated phrase.*

Text patterns created from a concordance analysis can be annotated and used in several ways. Possible syntactical text patterns[1] such as "[`Agent`] V-active [`Patient` Action-NN] " can be expressed in an even more general semantic form, often corresponding to more than one text pattern, depending on the task. This is *shallow parsing*, based on shallow semantic parsing and ontology-driven information extraction [6] and implies that only part of the sentences are parsed, namely those with the required semantics.

An example of how the semantic parsing output of a textual patterns can be represented is the language ONTOLOG [100], section 3.1.4, (example *regulation[AGT:substance,PTN:process]*). In the SIABO project, the semantic parsing leads to a reduction of paraphrases into one semantic frame or concept feature structure in ONTOLOG. The structure is often called predicate-argument structure, where the predicate here is the verb regulates and the argument types constraints can be the semantic types *continuant* and *process* (example of such a structure is <*insulin, regulates, glucose transport*>).[2] In section 4.4.2, the constraints of the argument for a given relation in a shallow parsing manner is exemplified by *regulation[AGT:substance,PTN:process]*: $regulates_{sp}$.

The semantics of the reduced form, often called a type constrained relationship, can be mapped to several text patterns that have the same meaning. This can be used in semantic indexing, since the ONTOLOG-form can serve as an index that can be mapped to several patterns within both corpus and queries in an information retrieval-context [9, 8].

Additionally, in the work of [136] (section 4.4.2) and [138], the work presented in section 5.4.4, a framework for type constraints of relata on semantic relations reflects more syntactical forms of phrases representing a regulatory relationship.

Our aim is to create an analysis of selected verbs concerning *regulation*, *negative regulation* and *positive regulation* within a comparable frame of textual knowledge patterns similar to [8] and [41] as well as a more formal semantic developed on basis of [136], thus combining formal semantic analysis with a semantic annotation.

Concordance analysis is employed for analyzing verbs that represent regulation in the biomedical domain context. The context should help to identify ontological types of relata mapped into the ontology.

[1] These "syntactical text patterns" are sometimes abbreviated "text patterns" or "patterns" and does specify semantic roles, but to avoid confusing these with the semantic frames-parts we only use these notion.

[2] In predicate-argument triples the actual appearences should be present and not the constraints.

The next sections focus on the quantitative findings of the corpus analysis. Subsequently, semantic frames for regulatory events are categorized with respect to the biomedical domain for the most commonly used regulatory verbs. This results in the frame ontology in figure 5.8.

5.4.1 Annotation of concordances

In order to identify the usage of *regulation*, *negative regulation* and *positive regulation*, a concordance of all occurrences of a selection of regulatory verbs in a corpus consisting of 40,000 arbitrary representative PubMed abstracts (632,316 sentences) was created.[3]

The search, which was published in [138] covered the active singular verb forms of the six verbs: "regulates" (323 occurrences, denoting *regulation*), "inhibits" (781 occurrences, denoting *negative regulation*), "reduces" (699 occurrences, denoting *negative regulation*), "decreases" (1119 occurrences, denoting *negative regulation*), "increases" (3171 occurrences, denoting *positive regulation*), and "stimulates" (372 occurrences, denoting *positive regulation*) as well as the singularly active form of 21 other verbs with a similar semantics.

In this work, the BFO concept *continuant* was replaced with the Semantic Network concept *substance* (explained in section 5.4.4).

In total, 2,000 concordances on 20 regulatory verbs have been analysed. This was an attempt to investigate whether the <*continuant regulates process*> triple structure had the most widespread usage in biomedical texts, to discover as many syntactical text patterns as possible, and also to investigate differences and similarities in the usage of verbs with similar semantics.

This was accomplished by annotating whether a sentence containing a trigger verb had the contexts, "substances regulate processes"(*sp*), "substances regulate substances"(*ss*), "proccesses regulates processes"(*pp*), and "processes regulates substances"(*ps*). When the domain of the phrase was not molecular regulation, it was annotated *nondom* and if the verb was not used transitively, i.e. having only one argument, it was annotated *nt*. If the left hand context was unknown, for example if the information on the first argument was expressed in a sentence before, the denotation was *?p* or *?s* (notation overview is given in table 5.3). A table with the results and statistics, as well as a short annotation guideline, can be found at *www.ruc.dk/~sz/Regrel/thesis*.

5.4.2 Quantitative findings

At first, a few quantitative findings based on the simple semantic annotation are presented. The most basic measurable findings with focus on the semantics are the following:

[3] http://www.nlm.nih.gov/bsd/sample_records_avail.html, 2009

Table 5.2: Notation and glossary on textual annotation patterns.

Name	**Description**	**Example**
Instantiation of a text	*Natural language phrase*	ethanol inhibits 3h-gaba release
POS-tagged text or Syntactical text pattern??	*Grammatical syntactical tagging of a natural language phrase*	ethanol/NN inhibits/VBZ 3h-gaba/NN release/NN
Frame part/patterns or Surface text pattern* Syntactical text pattern??	*Syntactical patterns with additionally semantic roles*	[Agent] V-active [Patient Action/NN]
Lexico-semantic pattern*	*Syntactical patterns additionally tagged with semantic roles and semantic type constraints*	[Agent] V-active [Patient Action/NN] sp**
Semantically anotated text	*Natural language phrase tagged syntactical, with semantic roles and semantic type constraints*	AGT:[ethanol/NN] inhibits/VBZ PTN:[3h-gaba/NN release/VB] sp**

* Both *Frame patterns* and *Lexico-semantic patterns* are considered *knowledge patterns* as defined in section 1.1.2.

** Constraining types from Semantic Network, *substance* and *process*, detailed description can be found in table 5.3.

Table 5.3: Annotation of concordances

Meaning	**Annotatation**	**Overall count**
substances regulate processes	*sp*	757
substances regulate processes	*ss*	322
proccesses regulates processes	*pp*	383
processes regulates substances	*ps*	137
domain of concordance is not molecular regulation	*nondom*	269
non transitive usage of verb	*nt*	-
left hand context is unknown	*?p/?s*	206/88
right hand context is unknown	*p?/s?*	-
context was not any of the suggested semantics	??	(269)

- *sp* was the overall most dominant semantic annotation (supporting the proposition of section 4.4). Then came *pp* (general process descriptions or cases which were expressed as "expression of X regulates process Y") and *ss* (typical enzyme modifications and DNA-modifications) with 23 % and 18% respectively, and finally *ps* with only 7% of the cases as can be found in figure 5.7.

- Regulates frame(regulate + affects) has in common the distribution: 60% *sp* and 25% *pp*, 10% *ss* and 1% *ps*.

- Enzyme verbs (activates, elevates, inactivates, blocks and partly stimulates and inhibits): *ss* (and *sp*) preference)

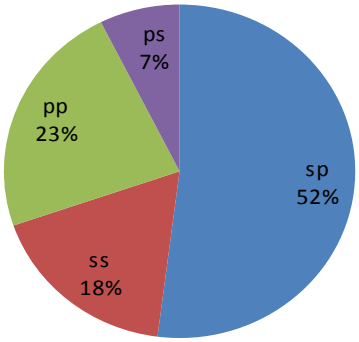
General linguistic observations shows:

- Many non-domain usages of affect, amplify, starts, limits, reduces (having a domain ratio under 80%). Some of these were, in the biological domain but did not express knowledge on molecular regulation. These verbs are also special cases, in the sense that they are very frequent in the BNC (especially start, reduce and affects, see, table 5.1) and might introduce noise in machine learning and semantic annotation.

- Verbs that express a different semantics (polysemi) (amplifies, generates, produces and removes) were generally high in *ps* and *pp* (method amplifies gene, process generates/produces/removes substance/process).

- The singular form was generally higher in *sp* and plural forms higher in *?s* and *?p*.

Figure 5.7: Distribution of semantic types based on 2000 concordances on 20 regulatory verbs.



5.4.3 Syntactical text patterns

The concordance analysis is utilized as a means of identifying the lexico-syntactic patterns that exist for regulatory verbs and their arguments in biomedical texts.

In this section, the analysis is concerned with the syntactical patterns found in concordances in the biomedical texts. These analyses consolidate the basis for a semantic extension/domain-extension of the semantic frames as presented in section 5.4.4.

Categorization of regulation relations

In addition to the quantitative findings, a deeper analysis of the concordance patterns will be presented. Through this analysis, four general types of regulations patterns have been identified, as outlined below and described in [138].

In this thesis, the concordance analysis has been extended to all 20 verbs, and five verbs in two different forms are deeper investigated. A fragment is shown in appendix B and the base for the statistics can be found at *www.ruc.dk/~sz/Regrel/thesis*. The quantitative differences in the verbs are presented in section 5.4.2.

By examining the concordances for these verb forms, the usage of the examined verbs, with respect to types of arguments, can be classified into four general frame types or patterns (analyzed further in section 5.4.4). 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.

The majority of the identified frame parts overlap with those identified in [27], however, two additional frame parts were identified through this analysis (*italicized*). The notation form for the frame parts presented below, along with some of the textual examples, are equal to the one used in [27].

- **Substances regulate processes**.
  This pattern covers roughly 50 percent of the occurrences of the examined verbs. Examples: "...ethanol inhibits 3h-gaba release...", "...glp-1 inhibits glucagon release...". This correlates with the frame parts in [27] having the syntactical text patterns:

  [`Agent`] V-active [`Patient` Action-NN]
  " IclR also represses the expression of iclR"

  [`Agent`] V-active [`Patient` *'production/ secretion/transcription/ expression/synthesis/ release'*] (added as an extension)

  [`Agent`] V-active [*'(the) synthesis/production/ secretion/expression/ transcription/ release of'* `Patient`] (added as an extension)

  [`Agent`] V-active [`Patient` *'activity/function'*] (added as an extension)

  [`Agent`] V-active [*'activity/function of''* `Patient`] (added as an extension)

  [`Agent`] *'is required/essential/involved in'* [`Patient` Action-NN]
  "rpoS function is essential for bgl silencing"

  [`Patient` Action-NN] V-passive [`Agent`]
  "yeiL expression is positively activated by Lrp"

  [Patient Action-NN] V-passive (*'caused by'*) [Agent]
  "bgl silencing caused by C-terminally truncated H-NS"

  [Action-NN 'of' `Patient`] *'by'* [`Agent`]
  "transcription repression of the Escherichia coli acetate operon by IclR"

  [`Patient` Action-NN] V-active [`Agent`]
  "Expression of the tau and ssu genes requires the LysR-type transcriptional regulatory proteins CysB and Cbl"

  [`Patient` Action-NN] V-active (*'caused by'*) [`Agent`]

- **Substances regulate substances**.
  This pattern covers roughly 20 percent of the occurrences of the examined verbs. For this pattern, the regulated substances are most frequently enzymes. Examples: "...lithium inhibits the enzyme glycogen synthase kinase-3...", "...rapamycin inhibits the kinase mTOR...". In terms of [27], the frame parts would be:

  [Agent]-Action-JJ [Patient]
  "SlyA-induced proteins"

  [Agent] V-active [Patient] (added)

- **Processes regulate processes**.
  This pattern covers roughly 25 percent of the occurrences of the examined verbs. Examples: "...delta and mu opioid receptor activation inhibits spontaneous gaba release...", "...nitric oxide pathway regulates pulmonary vascular tone...". In terms of [27], the frame parts for this pattern would be:

  [Agent Action-NN] V-active ('*cause*') [Patient Action-NN].
  "Disruption of cueR caused loss of copA expression"

  [Agent Action-NN] V-active [Patient Action-NN]
  "Elevation of ppGpp levels in growing cells... triggered the induction of all usp genes"

- **Processes regulate substances**.
  This pattern covers a minor part (5 percent) of the occurrences 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...".

Therefore, not many textual instances of regulations have a process on their left hand side, i.e. not many present the patterns *processes regulate substances* or *processes regulate processes*. But, the vast majority of the examples present a pattern where a substance regulates a substance/process; normally, when regulation relations are represented in biochemical interaction webs such as KEGG [73], 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". This leads to the extension of patterns in the next subsection.

The over-representation of the pattern *substances regulate processes* is also reflected in the number of text patterns found. For example, in [27],

thirteen patterns are found of which at least seven represent the form, *substances regulate processes*, two represent the form *substances regulate enzymes/proteins*, two represent the form *processes regulate processes*, and two have been difficult to categorize. A semantic discussion of this is given in section 5.4.4 and is also discussed in section 4.4.2, and in the reasoning part of the discussion chapter, section 7.3.

In an extended corpus analysis, as well as frame-comparison, from the resources of FrameNet, VerbNet and WordNet, it was observed that some of the verbs representing regulatory relations exhibit "deviant behavior", in comparison to the identified frame parts. For example, the verbs "increase" and "decrease" often appear in a passive or nominal form and, in these cases, the verbs do not have an expressed agent. Therefore, frame parts have been added, 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.

5.4.4 Relata and constrained relations

This section, will move from the representation of relations as text patterns, to top-level semantic frames with expressions like those in section 4.4.2. A small ontology of regulates is presented, with the purpose of categorizing different kinds of regulations as they occur in texts with respect to their lexical-semantic frames (figure 5.8). This distinguishes the ontology from other regulation-ontologies such as [21, 139], which are focused on the intensional meanings of concepts or types.

The analysis of the possible transformation of regulates text patterns has been described, along with an account of the semantics of these relations and a discussion of the types of relata. Though the proposed transformations are purely formal, they can be useful for a reasoning process, as well as for a foundation in semantic annotation.

The results of the corpus analysis as presented in section 5.4.3, can be viewed as an extensional definition of regulates 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 including the types or arguments attached to the relation (relata).

Types from Semantic Network

In line with [119], as presented in section 4.4.2, ontological types are distinguished from a top-level ontology. However, in this corpus analytic work, types from the domain specific top-level ontology of UMLS, the Semantic Network [92] are used instead. By using the Semantic Network as the top-ontology, it is possible to identify the ontological types of terms present in

the text.

Practically speaking, this can be done using the domain specific semantic entity-tagger, Metamap [12, 13], which links tokens to the UMLS as well as Semantic Network.

A similar attempt was made using the BFO as top ontology for annotating 97 full articles. However, the linkage between domain specific ontologies in OBO and the BFO was found to be lacking as was found in [19].

This means that the aforementioned knowledge patterns can be processed so that the semantic constraint, *substances inhibit processes*, can be incorporated into the knowledge patterns using concepts from the Semantic Network. Some examples of this will be discussed later in section 5.4.6.

As examples of the ontological types that restricts the semantic roles, "Substance(T167)" is a type with subtypes such as "Amino Acid Peptide or Protein", "Enzyme" and "Chemical." Additionally, "Phenomenon or Process(T067)" is a type representing "process" with sub-events such as "Physiologic Function" and "Cell-function."

Since all concepts in the individual UMLS resources have a direct link into the Semantic Network, this method makes it possible to capture the ontological types of a large number of domain-specific terms.

Substances are similar to, for instance, "continuants" in BFO [35], entities that continue to exists over time that 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 the relation *regulates*, there are four possible combinatorial relations among individuals, combining the two types of relata *substance* and *process*, corresponding to the four general patterns given in section 5.4.3. These relations are named $\mathbf{regulates_{ss}}$, $\mathbf{regulates_{sp}}$, $\mathbf{regulates_{ps}}$, and $\mathbf{regulates_{pp}}$, where the subscript "*ss*" means that the relationship can only exist between two substances; "*sp*" means that the relationship can only exist between a substance and a process; "*ps*" means that the relationship can only exist between a process and a substance; and, finally, "*pp*" means that the relation can only exist between two processes given in section 5.4.3.

Four general types of patterns discussed in section 5.4.3. $s, s_1, s_2, \ldots$ can thus be formalized 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 an *output_ of()* operator as proposed in [136], makes it possible to reduce these four relations to one, namely **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 **regulates$_{ss}$**, **regulates$_{ps}$**, **regulates$_{pp}$** and **regulates$_{sp}$** can be transformed into one, namely the **regulates$_{sp}$** relation. These transformations are given below:

s_1 **regulates$_{ss}$** $s_2 \Rightarrow$ s_1 **regulates$_{sp}$** $production\_of(s_2)$

s **regulates$_{sp}$** $p \Rightarrow$ s **regulates$_{sp}$** p

p **regulates$_{ps}$** $s \Rightarrow$ $output\_of(p)$ **regulates$_{sp}$** $production\_of(s)$

p_1 **regulates$_{pp}$** $p_2 \Rightarrow$ $output\_of(p_1)$ **regulates$_{sp}$** p_2

Additionally, a pattern denoting a slightly different meaning is noted to frequently occurs:

$$s_1 \textbf{ regulates}_{\textbf{sp}}\ function\_of(s_2)$$

This pattern denotes a regulation by a substance of the function of an enzyme or another substance and is thus within the sub-domain of enzyme kinetics.

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*:[4] "Insulin **stimulates$_{ss}$** glycogen," "insulin **stimulates$_{sp}$** the glycogenesis," and

[4]This example is also used in [136].

"insulin **stimulates$_{sp}$** the production of glycogen (through the Glyconeogenesis),"[5] where the process glycogenesis is equal to "the production of glycogen." Likewise, the sentence: "beta cell secretion **stimulates$_{pp}$** glycogenese" can be constructed and transformed to "output of beta cell secretion **stimulates$_{sp}$** production of glycogen," where the output of beta cell secretion is insulin.

A deeper discussion of the process of glycogenesis shows some of the implications of the **stimulates$_{pp}$**: Glycogen is an output of this process, but other outputs occur as well, for example uridine diphosphate (UDP), whose effects might be different from that of glycogen. Thus, when a statement that the glycogenesis stimulates glucose homeostasis is presented, it is not certain whether glycogen, or UDP or both are the actors, unless this is stated explicitly. Nevertheless, it is either glycogen or UDP (or both) that stimulate homeostasis, and not actually glycogenesis.

5.4.5 BioFrames ontology

This section moves from the representation of patterns in texts to the top-level semantic frames. A small ontology of regulates, with the purpose of categorizing different subtypes of regulations extensionally as they are present in texts, is presented. It, therefore, differs from intensional regulates-ontologies as [21, 139]. Characteristics of some of the bioframes have been made in a FrameNet-like manner, which can be found in appendix C.

The frame *BFN.Regulation* Regulation is an original suggestion and is in relation similar to the Semantic Network's *affects* relation. Notice that there is not a 1:1 correspondence to the frame of a verb and the semantic distributional characteristics as found in section 5.4.2.

Section 5.4.3 analyzed the extensional representation of regulation relations. However, to be able to annotate semantically, there is a need for an understanding of the intensions with the relations. Ontologies/terminologies have been developed in [21] and [139] concerning regulation as concepts.

OWL implementation

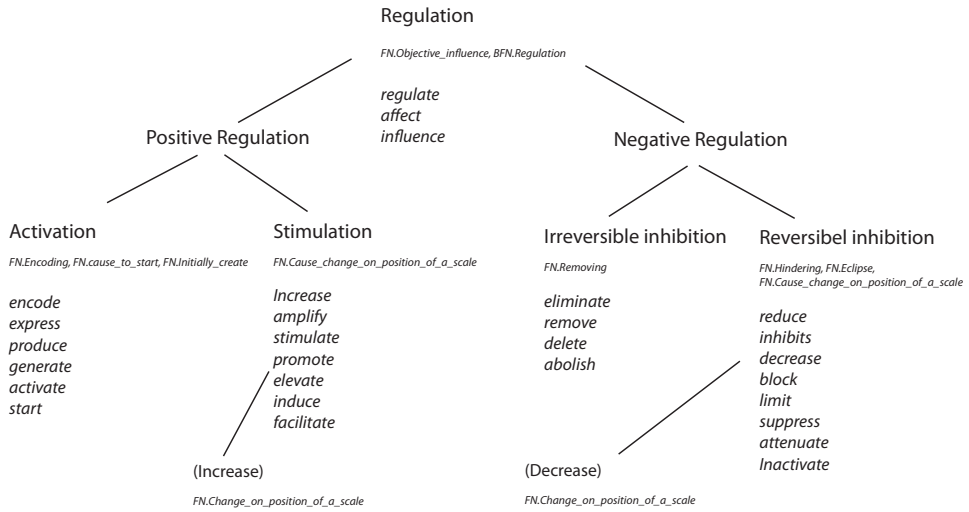
The ontology based on the BioFrameNet-style [42] is also modeled in OWL. In contrast to most other OWL-ontologies, which are focused on classes, this ontology has only two classes aligned with The Semantic Network and a hierarchy of relations, called *object properties*. Using the annotation specifications for each object property a description for each relation was created, explaining the formal definition or the semantic frame for the verb, if available. Additionally, corresponding terms are filled out for each property using language annotation.

[5] Note that *glycogenesis* and the above mentioned *glyco**neo**genesis* are two different processes.

Table 5.4: BioFrames and their corresponding verbs. Super frame inherits lexical units from sub frames.

BioFrame	**Lexical Units**	**Frame Role Elements**	**Super Frame**
Regulation/ cause change of position on a scale	regulates.v, affects.v, influences.v, (modulates.v?)		(Event)
Positive_regulation	*Only inherited*		Regulation
Negative_regulation	*Only inherited*		Regulation
Cause_to_start	start.v, activates.v, generates.v, produces.v, encodes.v,	Agent, Patient	Positive_regulation
Stimulation	stimulates.v, increases.v, promotes.v, elevates.v, induces.v	Agent, Patient	Positive_regulation
Increase	increases.v, increase.n, stimulation.n, elevation.n, promotion.n	Patient, (Agent)	Positive_regulation, stimulates
Cause_to_stop	removes.v, eliminates.v, abolishes.v, deletes.v	Agent, Patient	Negative_regulation
Hindering	reduces.v, inhibits.v, blocks.v, suppresses.v, limits.v, inactivates.v, decreases.v	Agent, Patient	Negative_regulation
Decrease	decreases.v, decrease.n, inhibition.n, reduction.n, blocking.n, suppression.n, limiting.n, inactivating.n	Patient, (Agent)	Negative_regulation Hinder

Figure 5.8: A verb frame ontology based on FrameNet, WordNet and our own corpus analysis. The frames are listed in table 5.4 and the verbs are treated in a frame analysis in appendix C [138].



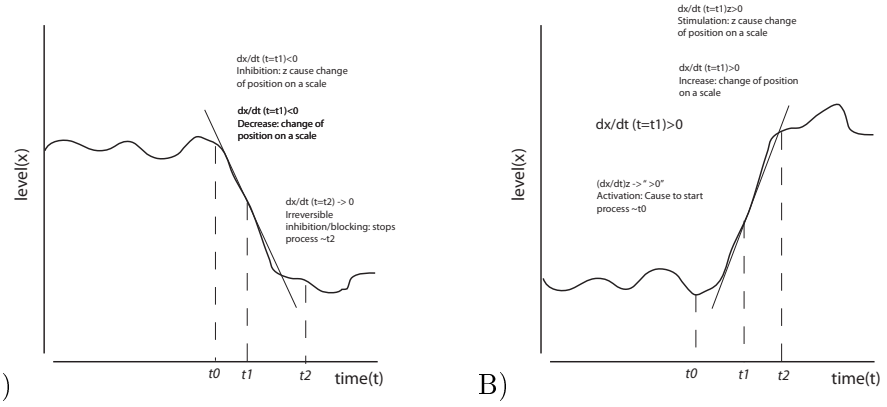
This ontology can be used within other ontologies if the modeler needs these relations as well as an extensive list of terms corresponding to the relation. The Regulation Frame Ontology is structured similar to e.g. Biorel [23] and can be found at *www.ruc.dk/~sz/Regrel/thesis*.

Kinetic description of bioframes

Earlier, in figure 4.2, the concepts of positively and negatively regulates was described by mean of a curve based on the level of x as a function of time, t. The different forms of positive and negative regulations identified in this section can also partly be described with with similar graphs as illustrated in figure 5.9.

Activation and irreversible inhibition can be described as punctual events well-defined on the time axis (corresponding to t_0 in figure B and t_2 in figure A), and stimulation and reversible inhibition are approximately events, which could happen anywhere on the time axis. The frame "Cause change on position of a scale" corresponds to the latter and, for example, occurs in t_1 in both figures. The verbs corresponding to irreversible inhibition are more complex, though, since they contain the information that "in all future, the molecules cannot gain the same amplitude as before the negative regulation."

Figure 5.9: The level of x as a function of time, t. A quantitative way of illustrating the enzyme kinetic forms of positive and negative regulations corresponding to the lexical frames in figure 5.8 and table 5.4.



5.4.6 Lexico-semantic patterns

As part of the aim of this work 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 they are known to mean. In this cross-field, between form and meaning, it may be possible to grab meaningful contents from texts through the semantic roles (e.g. agent and patient) of the relata.

To return to the types of the Semantic Network and connect those to the text patterns, the textual patterns will now be linked with the semantic relations and their types within the Semantic Network.

For example, s **regulates$_{\mathbf{sp}}$** p can have the following pattern, transforming traditional semantic roles into more constrained types from the upper-level ontology, the Semantic Network:

$$\begin{aligned} &[\texttt{Agent}]\ V-active\ [\texttt{Patient}\ Action-NN] \\ \Rightarrow\ &[\texttt{Substance}]\ regulates\ [\texttt{Phenomenon or Process}\ Action-NN] \end{aligned} \tag{5.1}$$

Meaning that *Substance,* or any subtypes of substance within the UMLS resources, are allowed (including the nouns representing those types), *regulates* or any subtype relation (like *inhibits* or *stimulates* and the verbs representing those relations) are allowed and *Phenomenon or Process* (including the phrases representing those types) are allowed. Patterns that fulfill these requirements can then be said to represent the semantic triple, s **regulates$_{\mathbf{sp}}$** p.

Another example, mostly typical for enzyme chemistry could be:

$$\begin{aligned} &[\texttt{Agent}]\ V-active\ [\texttt{Patient}\ function] \\ \Rightarrow\ &[\texttt{Substance}]\ regulates\ [\texttt{Substance}\ function], \end{aligned} \tag{5.2}$$

meaning similarly that *Substance,* or any subtypes of substance, within the UMLS resources are allowed (including the nouns representing those types) and *regulates* or any subtype relation (like inhibits or stimulates and the verbs representing those relations) is allowed. This is formulated more formal in the pattern: s_1 **regulates$_{\mathbf{sp}}$** $function\_of(s_2)$. However, contrary to equation (5.1), the second relata is a *Substance* followed by an adjective like *function* or adjectives with a similar semantics, which will be analyzed further in next subsection.

Additional patterns on regulates$_{\mathbf{sp}}$

During the corpus analysis, specific lexical patterns expressing modification of the patient were identified, which can be of use for extraction of knowledge.

The pattern s_1 **regulates$_{\mathbf{sp}}$** $production\_of(s_2)$, corresponds to the more specific frame parts like:

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

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

that contains information on the patient process.

Similarly, the *function_ of* operator in s_1 **regulates$_{\mathbf{sp}}$** $function\_of(s_2)$, has the expressions:

$[Agent]V - active$[Patient*activity/function*]

[Agent]$V - active$[*activity/function of* Patient].

These text patterns provide the possibility of identifying the substance that is part of the patient argument although the patient is a process.

A last example of more specific patterns concerns the usage of the bio-frame "Regulation". Although the verb "regulates" has been categorized as denoting a neutral regulates relation, this is not always the case. In a number of cases, a pre-modification of the verb by e.g. the adverbs "negatively," "positively," or "down" changes the relation to the more specific *inhibits* or *stimulates*. The same counts for affects, generating the following syntactical text patterns:

[Agent] *negatively/down V-active(*Regulation*)...*

[Agent] *positively/up V-active(*Regulation*)...*

These semantic specifications are important for future system development, with the aim of extracting interactions on a molecular level as it is discussed in section 7.3.2.

5.5 Summary

In this chapter, linguistic methods are used to analyze regulation. Terminological modeling principles were introduced, building domain ontologies for two micro-domains, enzyme inhibition (as in paper [38] and [139]) and regulations in the insulin-pathway domain from [9]. Additionally, the OWL implementations of the ontologies in section 5 and 5.4 were presented.

Corpus analysis is performed to give insight into different ways that regulatory events are presented in biomedical texts. First, a quantitative method (frequency list) is used to rank the most frequent regulatory verbs in a Medline corpus and a corpus consisting of biomedical patents compared with the general British National Corpus, BNC. These frequencies indicated that verbs involved in description of regulatory events are both overrepresented in biomedical patents and biomedical articles and are analyzed in [133] and [134].

Next, a concordance analysis was performed, in which lexical text patterns surrounding the regulative trigger verbs were identified. These were transformed into a semantic pattern that relates to that of section 4.4.2, described in a FrameNet-style, and included in a frame ontology of figure 5.8. The possible applications of this pattern extraction will be discussed further in chapter 7.

This work on lexico-semantic patterns is published in [138], though the section 5.4 contains a more thorough analysis, extended to all 28 verbs, as well as a further discussion on frames. Additionally, the regulation verb classes are illustrated in this dissertation by graphs based on regulation mechanisms.

Chapter 6

Semantics and retrieval applications

Receiving information and knowledge on regulatory events can serve many purposes. For example, information retrieval systems (IRS) can be based upon several reasoning rules, formalized ontologies on regulatory events, and semantic indexing.

The goal of an IRS is to retrieve documents containing information of interest. This process contains elements of preparation of the documents such as indexing, parsing of query and matching query to the text or information representation, as presented by [68].

A specific branch of information retrieval is what will be called *semantic information retrieval*. This branch covers mainly ontology based information retrieval (IR) and/or IR using semantic indexing based on KR, reasoning and linguistic models.

The SIABO project [9],[1] of which this thesis work is a part, will be presented in section 6.1. The acronym is an abbreviation for Semantic Information Access through Biomedical Ontologies. The main purpose is a semantic search that contains elements of semantic annotation and indexing as well as formal representation within biomedical ontologies and literature.

In this chapter an IRS developed in connection to the SIABO project (section 6.1.1) is presented, along with the conceptual work on semantic search in small biomedical corpus (section 6.1.2). These systems will be compared with other semantic information retrieval systems in section 1.2.4.

Additionally, some smaller prototypes that utilize the reasoning rules and the semantics defined in [134, 136] is presented in section 6.2. All ontologies are at a toy size, meant as a basis for specifications for the domain of molecular regulation.

[1] www.siabo.org

6.1 Ontology-based information retrieval

In contemporary information retrieval tools, the search is moving beyond keyword-based search. Many IRS products are domain-specific like PubMed or general like Google, have a strong bias towards fast systems in favor of the theoretically best methods.

In information retrieval, also vector space models, fuzzy systems and term-weight are important methods for intelligent query and answers. However, these will not be introduced in this thesis, except for a brief introduction of fuzzy systems in the discussion.

In ontology-based search, the knowledge of the semantics of the concepts in ontologies is utilized. For example, synonym terms, related terms, subtypes (query-expansion) and supertypes (object-expansion) can be incorporated into the system such that a query can be analyzed and mapped to the semantically closest documents.

An important aspect of this is the ability to deduce implicit fact based on the background knowledge. This corresponds to reasoning within the ontology as is treated thoroughly in this thesis in reasoning rules of section 4.5 and the compositions of $\mathcal{CRL}$, as described in section 3.2.

These ontology-based search techniques are especially strong within smaller domains. They can help to utilize biological domain knowledge in connection with the retrieval and thus make the search even more precise, though it requires extra computational power. A biomedical example on a sparse text corpus is presented in section 6.1.2, which is based on [62].

6.1.1 The SIABO project

The SIABO project is a project on ontology based information retrieval, partly based on the methods developed in the ontoquery project [98, 99]. Some of the methods are described in later papers: semantic indexing [8, 103], ontology modeling [88] and retrieval [9].

The aim of the SIABO project is to develop methods of querying by extracting knowledge from biomedical texts using biomedical ontologies. The motivation for this is the ever-increasing amount of research papers and patents from diverse resources within the field of biomedicine that challenge the process of retrieving relevant papers and information. To be competitive, biomedical companies need to have access to the contents of this increasing amount of documentation about their products, processes and projects.

The approach to this challenge is to develop methods that represent, organize, and access the conceptual content of biomedical texts using a formal ontology. The properties of an ontology-based system lead to easier access to data sources, locally as well as globally.

The SIABO approach introduces the notion of generative ontologies as described in section 3.1.4. The project sets up a novel, ontological semantics,

which maps the conceptual content of phrases into points in the generative ontology. Text chunks with identical meaning but different linguistic forms are mapped to the same node in the generative ontology. Thus, the approach facilitates identification of paraphrases, conceptual relationships and measurement of distances between key concepts in texts. The project focuses on ontological engineering of biomedical ontologies, applying lattices and relation-algebras, and has clear affinities to research in the Semantic Web area.

6.1.2 IR based on minimum text corpora

In addition to the broad focus of the SIABO project, an application on a minimum text corpus has been investigated within the domain of biomedical microarrays.

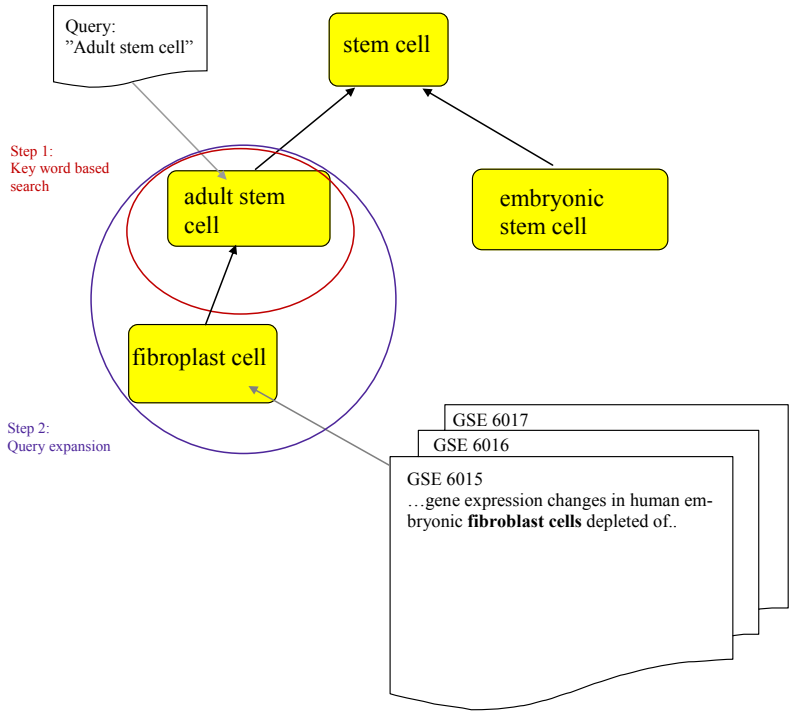
Gene expression profiles are kept on microarrays and used for data [33]. Though information exchange can be difficult due to the lack of standardization, a meta-data guideline does exist that outlines the Minimum Information About a Microarray Experiment (MIAME) [26]. The National Center for Biotechnology Information database, Gene Expression Omnibus (GEO), is a public functional genomics data repository supporting MIAME-compliant data submission [129], where it is possible to retrieve and download microarray data.

Retrieval of data is mainly based on a keyword text search, which is not always capable of finding all of the data of interest. Instead, the use of a bio-ontological oriented approach could be beneficial in such cases. However, creation of an ontological-based text search in connection with microarray experiments has only been investigated to a limited degree.

This subsection, based on [62], will focus on a more experimental approach to a (biomedical-)ontological oriented search. By demonstrating this technique through an example, the aim is to prove that it is possible to retrieve relevant information that otherwise would not have been found in an ordinary search. Also, this might lead to the development of a more intuitive approach to search for information in the microarrays. The biomedical data sources are based on meta-data from the GEO database, which was imported into our local database. By extracting information into an indexed text corpus, the thesis illustrates the potential to make a computerized text analysis using ontology. The method used to index the text corpus is similar to that in [7], which is closely connected to the SIABO project [9] and addresses problems of accessing the conceptual content of biomedical texts.

Similarly, the microarray-oriented MGED Ontology [129] uses the same approach to fetch data from GEO. They provide a framework that can be used by developers whose environment facilitates the usage of ontologies in microarray meta-data. It has not been convenient to use in this case, since it is difficult to extend with the current method of semantic language processing

Figure 6.1: The principles behind simple ontology based search. The query will be matched with the document by including sub-concepts in the taxonomy of the ontology.



[9] and since the focus in the ontology is on microarray techniques, rather than on content that describes, e.g. what the analyzed tissues consist of.

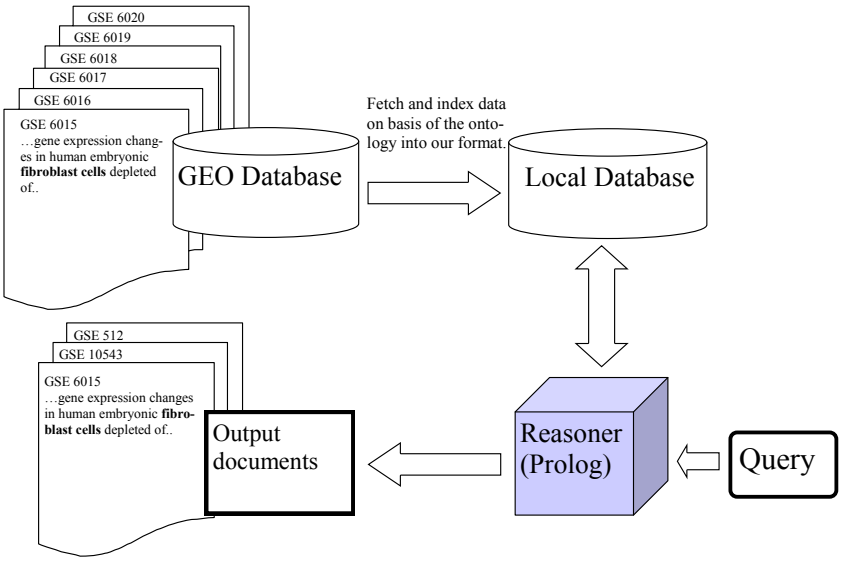
Methods and results

Gene expression experiments published in the GEO database are to be described in accordance with the MIAME standard [26]. Despite standardization efforts, the amount and quality of the information entered varies, and searches in databases can be quite a challenging task.

In order to demonstrate the possibility to improve the retrieval of microarray information, an example of a GEO experiment will be used. Prolog has been chosen for demonstration.

A GEO experiment is registered with a unique experiment ID in the database with links to microarrays, each registered with a unique microarray ID. An experiment is conducted on a specific platform and reuse of platforms

Figure 6.2: Illustration of the data flow from fetching the data from GEO to generating Prolog code and an end result [62].



occurs often. These platforms are also described, each with a unique platform ID, and can be cross-referenced. To support this ontology-based search method, a process fetches the data from the GEO database. The data is converted it into a new format and saved to a separate database. In this *process*, *entities* and properties of those entities, need to be extracted from the textual descriptions associated with the experiments.

To illustrate how this method works, a human stem cells experiment (GSE6015 [25]), which has been fetched from GEO is used. The experiment is recorded according to the MIAME standard and is well described in all aspects. It would be expected to find such an experiment through ordinary keyword search. Unfortunately, this is not the case with a keyword search for "adult stem cell," in which case GSE6015 will not be among the results. So-called "(embryonic) fibroblast cells," mentioned in the experiment text, are, in fact, adult stem cells. So, from a user perspective, this GSE is definitely a relevant result when searching for adult stem cells.

In figure 6.1, the principle behind ontology-based search is illustrated by simple queries and answers. By expanding the query, a document of interest is discovered, namely the GSE6015 containing the word "embryonic fibroblast cells."

The simple ontology in figure 6.1 is used to specialize a more general query "adult stem cell" to the more specific text "embryonic fibroblast cell" actually occurring in the textual description of the experiment. This scheme

illustrates just one form of expansion, since it is also possible to expand the query to more generalized concepts or through other relation edges.

6.2 Reasoning prototypes

A simple prototypes presented in this section is based on the reasoning rules of section 4.5.3. The Prolog regulation prototype is based on $\mathcal{CRL}$ and can be extended within a $\mathcal{CRL}$-context. Additionally, a probabilistic version has been developed using the language PRISM, though not published yet [95].

6.2.1 Prolog regulation prototype

In sections 4.4 and 4.5, a semantic and some reasoning rules connected to regulates as formal relation was presented. These should, of course, be evaluated to test their importance.

The most straightforward evaluation would obviously be an implementation of a prototype system for information retrieval and/or for hypothesis generation that would use the suggested formalization. A comparison with similar systems not using the same formal representation of regulatory relations, would then make the contribution of the semantic representation clear.

To illustrate the effects and properties of the relations constructed, a small example in the logical programming language Prolog was demonstrated. A small part of the KEGG database was implemented from figure 4.1 containing 21 classes and the relations: *is a*, *stimulates* and *inhibits*. Besides the relations in the figure, a small taxonomy was created to enable the separation of continuants such as (*small_molecule* and *gene_product*) and processes in correspondence to the way KEGG names the entities.

The toy-implementation can be used to infer fundamental inheritances in taxonomies of classes (ontology consisting of pure ISA-relations), as mentioned in the former subsection. This can be downloaded and tested from the website: *www.ruc.dk/~sz/Regrel/thesis*. Further work needs to be done to prove that the semantics of the implemented relations are actually equal to the semantics suggested in section 4.4.

6.3 Summary

This section deals with the feasibility, exploration and direct application of the dissertation. First, ontology-based information retrieval tools, such as the SIABO project are introduced. Next, [62] and the advantages of ontology-based search in micro-corpora 6.1.2 are presented.

Focusing on regulation, reasoning prototypes are developed in section 6.2.

Chapter 7

Discussion

Some of the chapters of this thesis are intrinsically connected. Thus, this chapter will discuss the works and points within three sections, focusing on knowledge, knowledge representation and reasoning.

This sectioning is chosen because these three levels represent, broadly, the main challenges in the work and since they are described in various ways with distinct methodologies in the different sections. First, to discuss the knowledge considered, next, to relate the representation of this in the different formalisms, such as $\mathcal{CRL}$, and, finally, to relate knowledge representation to reasoning and related applications. The purpose of this chapter is to both discuss the former chapters and to delve into the elements that have been particularly challenging.

A sketch will be drawn between what is represented by regulation, the means for formalizing relations and what is deducible by this relation (the work that is presented in chapter 4). The glue of the deductions over these relationships, class relationship compositions, is also discussed and examples of reasoning based on classes as is presented in chapter 3 are hypothesized. Last, but not least, a parallel to the linguistic-semantic work on how regulatory events are described in texts as in chapter 5, is included.

7.1 Knowledge on regulation

Knowledge, here understood as semantic information believed to be truly combined with reasoning possibilities, is the first-level to explore in this discussion. Knowledge on, e.g. regulation can be conceded in various ways, which is demonstrated in this thesis:

- *Ontology of regulation as a concept* is one way, which is performed by e.g. GRO [21] and by this author in [38, 139] (section 5.1.1).

- *Ontology of regulates as a formal (conceptual) relation*, can also provide knowledge as well as consideration of the relata of the relation and - as is possible for regulates - the kinetic-mathematics that underly this. This is represented in the section 4.2 and 4.4 and builds on the work in [136].

- *Corpus analysis*, looking into the verbs that represent regulates and how scientists use them is also an important source for defining *regulates*. This has been investigated in our publications [138, 133, 134] and is presented in chapter 5.

Whereas the two former hypotheses are mainly related to an intensional understanding of regulation, the latter is extensional in its ontological investigations. In the work on regulation, the understanding of the semantics of regulation became improved by complementary methods.

The frame-ontology of figure 5.8 was built upon the ontology of regulation in figure 5.2. However, after having examined the usage of the verbs, it appeared that, in addition to different usages, there were some differences in the semantics of the verbs. By comparing these linguistic differences with the mathematical understanding as shown in the graphs in figure 5.9, the ontology was revised. Thus, the frame ontology is an ontology with separating features reflecting usage and also has the perspective of meeting the classification of regulation as in e.g. GRO.

The knowledge of interest should always reflect the purpose of the system that is using it. A frame-ontology usually provides the purpose of reflecting written information and thus extracting knowledge. A conceptual ontology on the other hand, provides the purpose of clarifying and sharing the knowledge of the ontology, while formal semantics can help in reasoning over the relations. In information retrieval, the main requirement of the ontology is that it contains as many concepts/terms and variations over these within the domain of interest for providing extensional search.

These different purposes are reflected in the distinct ways of modeling throughout this thesis.

7.1.1 Usage of top-ontologies

During the work of this thesis, two of the top-level ontologies presented in section 2.3.1 were utilized.

For the ontological intensional work on defining the relation *regulates*, in the context of OBO and the Role Ontology, the OBO-recommended top-ontology BFO was used in section 4.4.2. The paradigm of this top ontology is realism, namely that all entities exist in the real world, as presented in section 2.2. This ontology is appropriate to relate the semantics of the relation to the other relations described, in terms of ontological types and entities from this ontology.

For semantic annotation and frame parts, on the other hand, Semantic Network, which is much less modeled in a theoretical science paradigm - if not what may be called pragmatic implementalism (in section 5.4.4) was used.

Semantic Network, which can be characterized as pragmatic implementalism, is, so far, the most convenient for this task since it operates with mapping to the resource ontologies and the semantic tagging tool Metamap was developed for this purpose.

Recently, a few studies have been based on the BFO for annotation, resulting in the CRAFT corpus [19]. So far, problems have occurred since "the OBOs have not been specifically developed for semantic annotation of natural-language biomedical documents" [19].

In future work, for comparison, both top-level ontologies should be tested on the same corpus for extraction of a combination of syntactic/domain text patterns as well as semantic types. Also, the more linguistically oriented ontologies, such as DOLCE and SUMO, should be considered for textual mappings.

7.1.2 Semantic patterns and text

As mentioned in the introduction of 7.1, the knowledge on a domain like regulation can be acquired in several ways. In section 4.4.2 some semantic frame parts on the form $\mathbf{regulates_{pc}}$, which have corresponding syntactical patterns in texts concerning the domain were identified.

The syntactic text patterns that are identified through the corpus analysis can form a background for further knowledge extraction using machine-learning. For example, text can be annotated automatically by use of the patterns and subsequently fed to a machine-learning algorithm for identification of new patterns (which is investigated in e.g. [60]). This automatic semantic annotation could create a basis for an ontology-based information retrieval. The semantic roles gained by such an attempt should be precise enough for inclusion in a specific knowledge base that could even be enriched with reasoning rules.

Additionally, through a deeper linguistic analysis, these parts can contribute to a domain specific FrameNet describing regulatory events, i.e. an extended BioFrameNet with domain specific semantic frames, in line with the vision of [42].

In section 4.4.2 a distinction between continuants and processes was made, leading to a characterization of four different basic relations among individuals. For stimulation these where the relations **regulates$_{cc}$**, **regulates$_{cp}$**, **regulates$_{pc}$**, and **regulates$_{pp}$**, which were further reduced to the single relation **regulates$_{cp}$**.

However, it may be argued that the relations **regulates$_{pc}$** and **regulates$_{pp}$** are not accurate relations in the first place; from a strict ontological point of view processes never stimulate other processes or continuants directly, but always stimulate processes through their outputs. This is also partly supported by the quantitative finings in section 5.4.2, illustrated in figure 5.7. The **regulates$_{cp}$** is clearly the most used relation in the semantically analyzed sentences.

On the other hand, in laboratory-situations, it is not always obvious which component in a process regulates, and thus **regulates$_{pp}$** (and **regulates$_{pc}$**) relations are used widely in biomedical abstracts. Another explanation of this is that describing a process as regulating another process is more general, even to biologists; the processes can be more familiar than particular molecules within each process.

Section 4.4.2 indicates: "The $production\_of(...)$ operator works on a continuant c by transforming it to the process that is the production of c. Similarly the $output\_of(...)$ operator transforms a process p to the continuant that is the output of p."

However, this is not necessarily that simple. For example, some processes may have several outputs which will not be specified alone by using the *output_ of()*-operator. In larger regulatory pathways, processes may regulate other processes, though always through outputs of the first process. These outputs can be unknown, or it can be unknown which of the outputs actually regulates the second process, which makes this **regulates$_{pp}$**-relation a bit imprecise.

An example of this is glycogenesis as described in section 5.4.4, glycogen is an output of this process, but other outputs occur as well, for example uridine diphosphate (UDP), where effects might be different that glycogen. Thus, in a statement that the glycogenesis stimulates glucose homeostasis, it cannot be certain whether glycogen or UDP or both are the actors, unless this is stated explicitly. Nevertheless, it is either glycogen or UDP (or both) that stimulate homeostasis and not actually glycogenesis.

It is a debatable issue whether processes can stimulate other processes or continuants. There seems to be evidence, however, that it is important to investigate the ontological aspects of stimulation further. Whether stimula-

tion is among continuants or processes seems to have consequences for the inference of new knowledge and, thus, the distinction should be recognized. In simple knowledge representations by graphs, such as the KEGG database, such observations are not accounted for. Such knowledge bases have the potential to get this representation integrated automatically when a semantic is agreed upon.

7.1.3 Corpus analysis and genre

Corpus analysis can be used as an appropriate tool to approve both the importance of the domain specific verbs and the underlying semantic relations. Biomedical verbs like *activate* and *inhibit* are more frequent in biomedical texts than in the BNC, and the verbs are, thus, worth investigating, since a distinct usage of the verbs is a possibility.

This low entropy of e.g. *activate* and *inhibit* indicates that the verbs contain more information than other, more common, verbs and that these could be added as representing relations to be utilized by a developer when building a background ontology. The relevance of the verbs could be benchmarked further using a weirdness test [59], for example.

A lot of the verbs used by e.g. Chilibot were not even present in any biomedical corpora, making them only marginally important, although they might play the same role as the more frequently used verbs.

It is not insignificant which text genre was analyzed. In the patents, though they only represent a narrow part of the biomedical area, some differences in biomedical word frequencies compared to Medline were still found. For example, the words *encode*, *inactivate* and *remove* were relatively frequent in the patents. One reason may be that the words are frequent in a legal context, which every patent contains alongside the domain-language of the claim. On the other hand, typical biomedical verbs like *increase*, *induce* and *decrease,* for example, were less frequent in the patent texts than in Medline as displayed in table 5.1.

Special cases from concordance analysis

In accordance with the analytic part of the corpus analysis, a few semantic-linguistic challenges occurred. The task was to unfold regulatory events, especially with respect to the agent and patient roles. However, besides the obvious and desired situation of a sentence like "insulin stimulates glucose transport," in many sentences the regulatory events were hidden and have many variations in expression.

For example, activate and stimulate, that apparently are very similar, also have an important difference. One had the frame *cause_to_start* and the other, *cause_change_on_position_of_a_scale*. In a regulatory chain, this is not necessarily important, but the little semantic difference might

have an influence. If trying to stimulate a molecule, e.g. a protein, that is not expressed/activated because of lack of a transcription factor or promoter there is no effect. Similarly, an inhibition of such a protein will have no effect since it is not activated anyway. However, for example, most of the metabolic pathways consist of gene products that are constantly produced and, thus, the assumption of a general positive and negative regulation is often adequate.

Another special case is non-transitive frames, referring back to a former sentence or an unknown factor. With sophisticated methods, these apparently non-transitive frames can occasionally be crystallized into a transitive relationship, but often the knowledge that is extracted is a regulatory relation and a patient. Although this contains knowledge in a whole abstract, it is difficult to extract knowledge on the agent of this process.

Finally, verbs like secretes, encodes and produces have a slightly different type of relata than other regulatory verbs focused on molecular interactions. These production verbs often contain types like *Cell* or *Organ* as the first argument and not *Substance* like "stimulates", for instance.

7.2 Knowledge representation

Knowledge representation (KR) is, of course, closely related with both background knowledge and reasoning over this knowledge. Before knowledge is acquired, it can be difficult to suggest an appropriate representation. The possible reasoning tasks that can be performed on the knowledge are limited by the representation as well, since very sophisticated rules might be prohibited by too simple a formalism.

This thesis has worked with different formalisms for representing knowledge, which all have their purposes; from the expressive formalisms, like first-order logic, to simple graph representations. This researcher's contributions within this areas lie mostly in the works [136, 4], as described in section 3.2 and 4.4.1.

Selecting the right formalism is a tradeoff between a KR that supports simple and tractable operations and one that is expressive and provides many opportunities for reasoning.

First-order logic($\mathcal{FOL}$) and description logics($\mathcal{DL}$) are both popular within the KR-field. Almost all biomedical ontologies are represented in the $\mathcal{DL}$-based OWL-format, but the tractability of the language is not very high. This might also be the reason that the most popular $\mathcal{DL}$-languages are light weight languages [17], which have only slightly more expressiveness than the tractable database language SQL, though still considering the relatively "computational expensive" open world assumption.

The formalization of vagueness and uncertainty should also be discussed when a knowledge representation is considered. Many formal relationships on classes have exceptions or uncertain instances that should be taken into consideration as is discussed in section 7.2.2.

7.2.1 Formal representation of regulates

The goal of section 4.4 has been to model the formal relationship as precisely as possible using $\mathcal{FOL}$. Although this representation language easily leads to intractable structures for practical purposes, it is always possible to simplify the language in the application.

The work on compositions of class-relationships (section 3.2) was originally carried out as a support to reasoning over the relationships of regulates defined as the $r_{\forall\forall}$ class-relationship. This reasoning topic will be discussed in section 7.3.1.

As can be seen in table 3.2, the $\mathcal{CRL}$ meta language is similar to $\mathcal{DL}$ and some of the relationships are also directly present in $\mathcal{DL}$ such as the $A\ r_{\forall\exists}\ B$ ($\tilde{}A \sqsubseteq \exists r.B$). However, the CRL framework is simpler, assumes closed world, and can easily be implemented in Datalog, for example. Additionally, it does not allow logical construct descriptions such as $\mathcal{DL}$, but only the binary relationships among classes based on predicate logic sentences using

quantifiers, which is adequate for many ontologies.

The advantage of using a $\mathcal{DL}$-language is that it is well-studied and used for several applications within the ontology and semantic web fields. In several of the $\mathcal{DL}$-flavours' reasoning is decidable, contrary to first-order logic, which is undecidable. $\mathcal{DL}$ has decidable fragments and many of the sub-languages provide reasoning using polynomial time and space. The $\mathcal{DL}$-based application, Protegé-OWL, is W3C's preferred ontology language and most biomedical ontologies are implemented in this language, as collected in e.g. [97] (among other formats).

However, $\mathcal{DL}$ is an expressive language and the expressiveness comes at a price. Usually, large knowledge bases are heavy in $\mathcal{DL}$ and, for example, experiments with implementing the $\mathcal{DL}$-based DanNet[1] containing approximately 300,000 concepts in Protegé-OWL failed the first time because of problems with the size.[2] Additionally, OWL implements class-triple in too-complex ways, as noted in [23] and are mostly developed for the purpose of defining individuals and individual relations which motivates a $\mathcal{CRL}$-focus.

These kinds of problems seldom occur in querying a SQL-database of that size and it is possible to predict that $\mathcal{CRL}$ would also be capable of managing a knowledge base of this size, due to its Datalog-meta-logic framework. Other works on, e.g. the OWL EL fragment, have also shown that implementing part of the ontologies as inference rules in Datalog increases the efficiency for, e.g. instance checking and classification [77].

7.2.2 Vagueness aspects of regulates

In section 4.4 a semantic analysis of regulates, in which the term "can potentially" plays a considerable role, was performed. It is a vague modal term that could be more precisely formulated, suggesting different understanding of the vagueness or uncertainty.

When representing "insulin stimulates glucose transport" as:

$$\forall x(Insulin(x) \rightarrow \forall y(Glucose\_transport(y) \rightarrow x \ \mathbf{stimulates_{cp}} \ y)),$$

the term "can potentially" is implicit in the relation "$\mathbf{stimulates_{cp}}$."[3] In other words, "x $\mathbf{stimulates_{cp}}$ y" is read as "x can potentially stimulate y."

A first vagueness is due the fact that stimulation only takes place if the substance is actively participating in the process. If the process and substance are separated in space and time, stimulation can of course not take place. The relation of a continuant taking active part in a process at a given time, is a basic relation and in [119] it is assumed as a primitive relation. Using their notation, "p **has_agent** c **at** t" expresses that the

[1] www.WordNet.dk

[2] Early experiments presented at the DanNet symposium 2009

[3] Note that the *cp* in $\mathbf{stimulates_{cp}}$ stands for the relata *continuant-relation-process* and not "can potentially"

continuant c is causally active in the process p at time t. Together with the **stim** relation, "x **stimulates**$_{\mathbf{cp}}$ y" this can thus be expanded as:

$$\forall t(p \textbf{ has\_agent } c \textbf{ at } t \rightarrow c \textbf{ stim } p),$$

A second vagueness that can be implicit in a relation is the uncertainty of predicting interactions. In non-trivial examples, such as the predicted regulation by miRNAs as described in section 4.1.1, the stimulation is only *predicted,* and not based on laboratory evidence. For example, $miRNA$ **stimulates**$_{\mathbf{cc}}$ c (predicted *in silico*), and c **stimulates**$_{\mathbf{cp}}$ p should lead to a weaker regulatory causal relationship between $miRNA$ and p than if the $miRNA$ was experimentally shown to stimulate c.

A third vagueness is based on the trustworthiness in the knowledge-extraction of the written word and corresponds to the strength of the knowledge as discussed in section 3.2.1. A canonical regulation in a database might differ from that described in texts. For example, a sentence like, "Our findings suggest that protein A down regulates Process B," could be a statement built on numerous experiments ($r_{\forall\exists}$ or $r_{\forall\forall}$), or it could be a stand-alone result, an $r_{\exists\exists}$-relation or even a controverse statement. In an implementation, this could be handled by automatically noting any regulates-relation "$r_{\exists\exists}$," and if it is supported by other findings (in other abstracts) it should be upgraded to a $r_{\forall\forall}$-relationship.

7.3 Reasoning over *regulates*

In this section, different functionalities of reasoning in prototypes within the frames of two different approaches of using knowledge on regulation are suggested.

First, reasoning on graph, or logically-based knowledge bases, is valuable for biologists. This could be a standard hypothesis-testing machine that will require some probabilistic uncertainty, since the causality will decrease through a pathway. The more subsidiary factors interfere with regulation, the smaller the possibility for the regulatory chain. This is discussed in next section.

A second approach is to extract knowledge from text, using semantic patterns and reasoning over the knowledge. In this case, the relationship-triples used for semantic indexing in information/knowledge retrieval systems are extracted. For this purpose, the semantic patterns in [138, 136], as well as the semantic frames can be of use as will be discussed in 7.3.2. This can link the regulatory relationships to the SIABO project presented in section 6.1.1.

7.3.1 Hypothesis support and graph reasoning

From the reasoning rules of section 4.5, and the composition rules in section 3.2, additional relationships can be deduced from existing ones to infer things like: "if insulin stimulates glucose transport and if the glucose transport inhibits glyconeogenesis, then insulin inhibits glyconeogenesis" (as in the prototype in section 6.2). Thus, if seeking novel gene products and molecules that regulate a given process or a given molecule in a certain way, reasoning rules can be used to predict these. Another aspect of this automated reasoning is the prediction of the side effects of a drug or extra, perhaps unknown, molecule functions.

Furthermore, a new, unfamiliar, molecule can be placed correctly in a regulatory pathway due to its regulatory properties. These functions can be an advantage in drug discovery, identification of adverse effects and in knowledge expansion for more fundamental research purposes. These possibilities are just some of the many advantages of a logic-based knowledge representation, as the one presented here, could provide when fully implemented.

Vague and probabilistic reasoning

The inferences are also characterized by vagueness as several occurrences in chapter 4 and section 7.2 reveals. As Laplace mentioned in 1814 (translated from the French version of 1825) [79]:

> "All our knowledge is problematical; the entire system of human knowledge is connected with the theory of probability"

This sentence eulogizes elegantly an underlying problem in understanding of ontologies as deterministic relationships. And, as already mentioned in section 7.2, regulates as formal relation has vagueness which is even more obvious than the one of formal isa-relationships.

This could be taken one step further by applying fuzzy logics, or other logics of uncertainty, to model this aspect of regulatory relationships. This could be in terms of linguistic variables or relationship types like $r_{\exists\exists}$ can $r_{\forall\forall}$to get a value assigned fuzzy sets [61].

In another possibility, with respect to the prototype of section 6.2, is an experiment with a version using the probabilistic prolog-plug-in, PRISM, with which it is possible to add probabilities to the relations as well as the relationships and chains [95]. A similar approach, in a dynamic language, is of course also possible. This would normally lead to faster programs, although declarative languages have the advantages of integrated inference machinery, and fit in their code structure with declarations in logics.

Transforming deterministic ontologies into stochastic ontologies brings ontological knowledge management within the realms of statistical modeling. Introducing probabilities in ontologies enables the use of the structure of ontologies, even in cases of missing data in terms of hidden relations. It also handles situations in which various individuals have distributed membership.

Probabilities are convenient because they both describe the frequency of distributed instances and form a normative rule for updating beliefs and degree of reliability of an event or relationship. Stochastic ontologies give a quantitative account of deterministic ontologies, to estimate the uncertainty inherent in any classification, and have potentials in e.g. regulation reasoning systems.

Complexity of $\mathcal{CRL}$-based representation

One of the main advantages of modeling knowledge, in a (quite expressive) formal framework as logic, is that it makes entire knowledge bases more complete and allows for the use of reasoning tools to gain new knowledge. This is particularly useful in, for instance, AI and semantic information retrieval.

In relation to the discussion of representation forms of regulation-networks of section 4.2, this logic, $\mathcal{FOL}$-formalization, is in the middle of a complexity scale. It is not as expressive as the linked differential equations [64], but much better suited for automatic reasoning than simple graphs [73]. In expressivity and tractability it is similar to work like [43, 30]. However, the logical analysis of *regulates* in section 4.4 provides a semantic and uses first-order logic formalization, which expresses more information to the relations than simple graphs or propositional logics, due to the quantifications.

Complex role inclusions using class relationships are not as obvious as complex role inclusions among individual relations. Through these compo-

sitions, it is possible to reason over the rules that the knowledge engineer defines for example over the reasoning rules of section 4.5. As an additional benefit, the compositions also provide possibilities for complex role inclusions among different class relationship-types.

An example of problem with the semantics from Gene Ontology, captured by $\mathcal{CRL}$-compositions is the following reasoning rule [51]:

$$regulates \circ partof \sqsubseteq regulates.$$

Considering classes, the composition would look like following table 3.3:

$$A\, regulates_{\forall\forall}\, B \;\odot\; B\, isPartOf_{\forall\exists}\, C \sqsubseteq A\, regulates_{\forall\exists}\, C.$$

Through this composition, regulates looses its $\forall\forall$-property. This would not occur with the opposite rule:

$$A\, isPartOf_{\forall\exists}\, B \;\odot\; B\, regulates_{\forall\forall}\, C \sqsubseteq A\, regulates_{\forall\forall}\, C).$$

This is a problem as soon as relationships become class relationships. These deductions can be made remembering to emphasize that regulates has lost its $\forall\forall$-property to a $\forall\exists$. This is a problem since new reasoning rules, presuming the $\forall\forall$-semantic for regulates cannot automatically be applied. A solution, however, is if $isPartOf_{\forall\exists}$ also has an inverse $hasPart_{\forall\exists}$, that is true for the situation (leading to constituating part-hood), the problem will not occur.

7.3.2 Semantic extractions for Knowledge Based Systems

This work on verbs and relations can be integrated into a larger knowledge model for use in domain areas in industries or academia. Three purposes of the formal semantic relations are suggested:

1. To model a background ontology for information retrieval, based on the semantics of the relations suggested.

2. To map different text corpora into the ontology by indexing concepts into biomedical text corpora and queries (on the fly) [9]. This is to fine-tune the output of the search, when the query contains regulation and, thus, to provide semantics within information retrieval tools.

3. To use automatic ontology generation that is trusted, based on the relations [111] and biomedical texts. When a few patterns are available, by using the relations as a hook, new concepts might be discovered.

Whether the identified verbs should be expressed simply, as relations, or added into the ontology as concepts themselves, is a question under discussion that the knowledge engineer should consider.

If the verbs are utilized for relations, they can be used for automatic ontology generation because the information can be used for capturing the surrounding semantic types. Additionally, they can be used as part of the similarity measure in an information retrieval context. Furthermore, because the verbs have transitive-like properties, they can be used to infer more knowledge than what is described explicitly in the sentences. Formal semantic relations ensure keeping as much information in the ontology as possible.

On the other hand, with as few relations as possible and a nominalization of the important verbs, will cause the number of concepts to grow; however, the resulting background ontology will be simpler as presented in [9].

Using the nominalized forms, a thorough indexing is possible, consisting of the verbs as part of the phrases, which can be easier to represent when matching queries or searches into the index. The particular purpose with the nominalized form is that it can act as a concept in an ontology instead of a relation and the nominalization usually collects the most semantic contributions of a sentence. Also GRO [21] uses nominalized forms of *regulate* in their ontology of regulation.

Semantic patterns and reasoning

In section 4.4.2, some basic regulates relations among individuals was introduced. On top of this, a textual correspondence to this was developed, introducing a semantic frame for different regulates semantics, as described in section 5.4.4 [138, 136].

This semantic foundation conveys a basis for extraction of regulates relations and could be the base for searching on papers with specific relations among potentially unknown molecules or exploring literature and the relationships among specific substances in an application.

For example, a structure of one sentence with many regulatory events like "the inhibition of B activates A." Another possibility is reasoning over a chain of regulations, such as:

$$S_0 - r - S_1 - r - S_2 - r - S_3 - r - S_4 - r - S_5 - r - S_6 ... r - S_n,$$

where $S_{0\rightarrow(n-1)}$ is either a substance or *production_ of(S)* or *function_ of(S)*.

Whether relationships revealed from texts are more than a simple $\exists\exists$-relationship and, thus, can be of any usage in reasoning (recall in table 3.3, that $R_{\exists\exists} \odot R_{\exists\exists}$ is not a valid composition) can be discussed. However, in a corpus like PubMed abstracts, what is written in this is the essence of a scientific investigation and thus parallels in strength of knowledge to database information, although the uncertainty is higher. A probabilistic approach, such as the one discussed in section 7.3.1, could be adequate, and even supply knowledge from databases with a relatively lower probability than those only deduced on the basis of the data bases.

Problems in the complex role inclusion and *sp*-semantics

In section 4.5, the concepts transitivity and inter-transitivity across relations like *inhibits* and *activates*, were introduced.

In reasoning, the transitivity and inter-transitivity functionality is more sensitive. A counter-example from the Gene Ontology is that, although regulations of antiapoptosis regulate the regulation of apoptosis, which regulates cell death, it cannot be concluded that regulation of antiapoptosis regulates cell death [51].

The question is whether or not this would count for regulations between substances and processes. For example, a substance could *regulate* the regulation of apoptosis, of which an output *regulates* cell death. Here, it is arguable that the substance indeed regulates cell death as well, supporting the reasoning/possibility of a chain of regulation-events, as described above. However, this might not be true for Gene Ontology, which is mainly concerned with processes rather than substances.

Another concern of the proposed reasoning is that the chain of reactions might be more insecure the more relation edges are traveled. Can anything really be said about a substance that regulates another substance in the fifth link? And should this be limited only by a lower probability or by a short-cut limit of the size of n in the chain?

This is a delicate issue and the example is not conclusive. However, care must be taken when using information to infer what is hoped to be exact knowledge. The transitivity and inter-transitivity of positive and negative regulation can be seen at as a rule of thumb, or a rule connected with some uncertainty leading to conclusions with a certain level of uncertainty. This is why the functionality is called "hypothesis support" rather than "hypothesis creation."

7.3.3 Micro corpus and information retrieval [62]

Small and shallow corpora as microarray experiment descriptions provide several challenges in reasoning in an information retrieval system. Recall that information retrieval basically handles the task of providing documents as response to a query. The content of a microarray experiment is mostly raw chip data, attached to the description, which is a small document containing information.

An important issue of the functionality of semantic search in microarray-corpora, as described in section 6.1.2, is that the texts attached to the chip data is annotated with differing qualities, although usually in line with the minimum requirements, MIAME [50]. Some, like the case of experiment GSE6015, have filled out most of the blanks with a lot of information, whereas others, as in GSE1310 (another embryonic stem cell study from GEO), lacks a lot of information. It is written according to the MIAME

standard, although very sparsely, and will be difficult to retrieve in a normal search.

Sparse annotations are a complication, inherent too much of the microarray-corpora, where an ontology-based approach may be advantageous, since it can provide a means to bridge the query to the experiment through ontological inference by expanding the query and queried text based on the ontology.

A similar problematic occurs in electronic patient records, where the symptom description is often very shallow and has an individual touch, depending on the physician writing it, just like the microarray experiment descriptions.

From the initial experiments, ontology-based search for microarray data is found to be potentially valuable and is an area that needs more investigation.

Summary of contributions

The work in this thesis is an assortment of published articles and additional contributions, rather than a collection of papers. Thus, an overview of the contributions of this author will be offered, along with a summary of research question conclusion. Details on sections and contributions will not be summed up.

Within the domain of biology, an attempt has been made to describe and explain general biology and prepare it as a foundation for a logical representation. This is regards collecting a set of illustrative regulatory examples, explaining the mechanism of regulation using appropriate simplified differential equations, modeling of ontologies on the subject and discussing the basic ontological assumptions of the domain.

Within the field of logics and computer science, a formal analysis of regulates was performed, as well as reasoning rules for a small implementation and a pilot work on compositions of class relationships has been developed.

Having a background as a biochemist has also allowed the researcher to play a role as domain-expert (or at least as a knowledge engineer quite familiar with the domain) in the SIABO project on semantic information retrieval and several knowledge acquisition works in this thesis

The main contributions of this thesis are, in summary:

a) The categorization of regulation as relation and its subtype-relations with respect to ontological types which is used both within a logic KR [136] and corpus analysis [138].

b) The suggestions of information systems that utilize a logic representation [134, 4].

In a sense, the main contribution can be seen as knowledge engineering of regulates relations within biomedical informatics. This is the background for the answer to the extended research question:

> *What is the semantics of regulation as relation within molecular biology and how can this be used for reasoning in hypothesis testing in a tractable way? How are regulatory events represented in biomedical texts and how can this information be utilized in the above mentioned problem?*

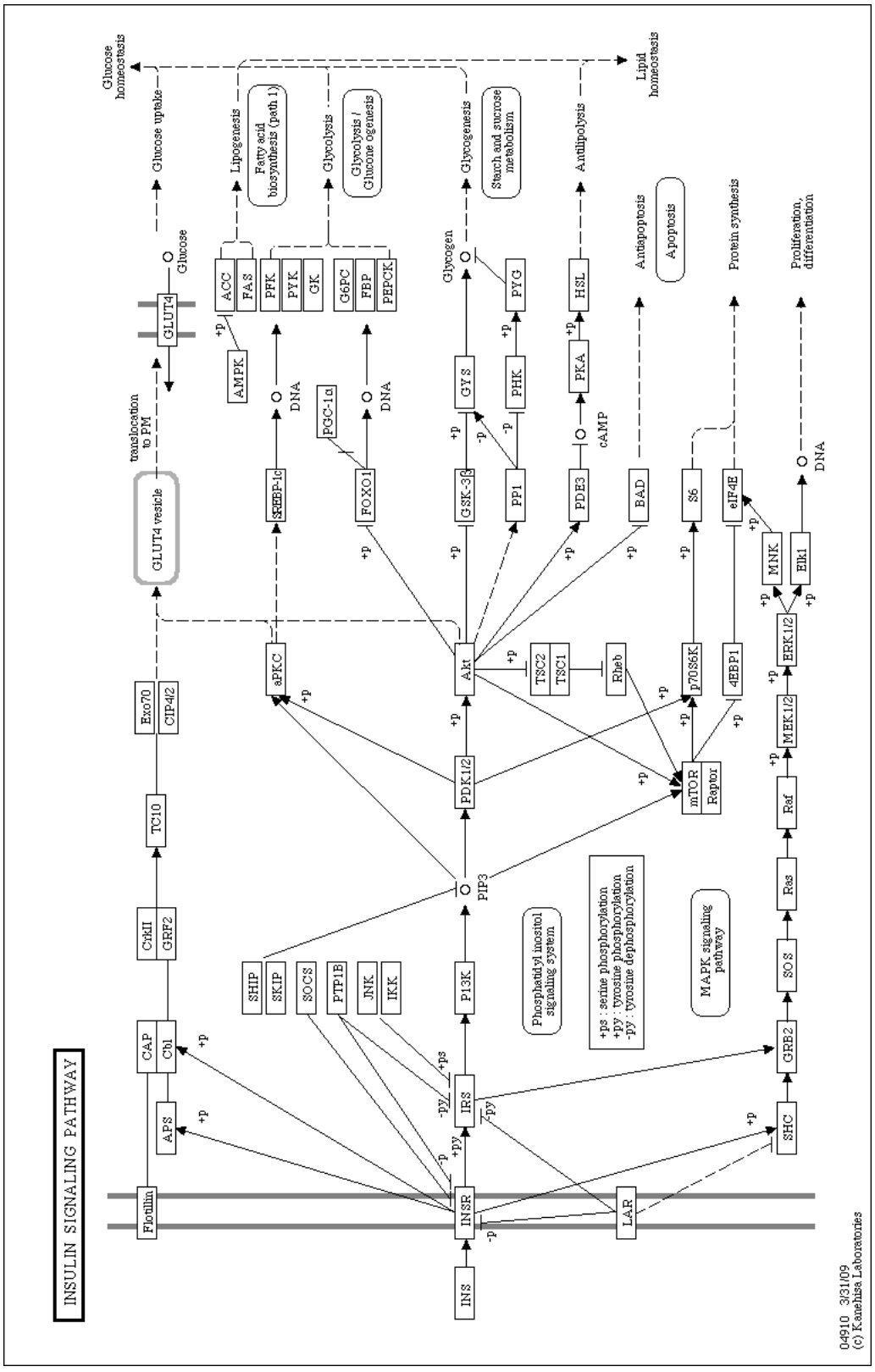
In chapter 4 (based on [136]), it was shown that the semantics of regulation as relation can be expressed intensionally by the $\mathcal{FOL}$-formular defining $\mathcal{CRL}$-relationships C_1 $regulates_{\forall\forall}$ C_2 iff $\forall x(C_1(x) \rightarrow \forall y(C_2(y) \rightarrow x$ **regulates** $y))$, cconsidering the individuals as "amounts." Additionally, regulates as relations are extended as a top concept in a regulates role hierarchy as in figure 5.8 (based on [138]), and also formalized later in OWL.

The opportunities in reasoning are suggested by presenting complex role inclusions for classes (defined in section 3.2, based on [4]) using the reasoning rules presented in section 4.5 (from [135, 134]). These rules should be implemented introducing an amount of uncertainties, e.g. by probabilistic reasoning or fuzzy logic within the chains of predicted reactions.

The appearances of regulatory events in biomedical texts were investigated by frequency analysis and concordance analysis. The distributions and results of this was displayed in sections 5.2-5.4 (based on [134] and[138]). This information can be utilized for extracting patterns that correspond to a few simple semantic frames (e.g $regulates_{sp}$ ~ $< substance - regulates - process >$), extracting the relation and the involved substances to build a system on these extracts as discussed in section 7.3.

Appendix A

Insulin signaling pathway from KEGG


INSULIN SIGNALING PATHWAY
Glucose homeostasis
Glucose uptake
Lipogenesis
Fatty acid biosynthesis (path.1)
Glycolysis
Glycolysis / Gluconeogenesis
Glycogenesis
Starch and sucrose metabolism
Antilipolysis
Lipid homeostasis
Antiapoptosis
Apoptosis
Protein synthesis
Proliferation, differentiation
Glucose
Glycogen
cAMP
DNA
translocation to PM
GLUT4 vesicle
GLUT4
ACC
FAS
PFK
PYK
GK
G6PC
FBP
PEPCK
PYG
HSL
PKA
AMPK
PGC-1α
SREBP-1c
FOXO1
GYS
PHK
GSK-3β
PP1
PDE3
BAD
S6
eIF4E
MNK
Elk1
ERK1/2
MEK1/2
Raf
Ras
SOS
GRB2
SHC
Exo70
CIP4/2
TC10
CrkII
GRF2
CAP
Cbl
APS
Flotillin
aPKC
Akt
TSC2
TSC1
Rheb
PDK1/2
mTOR
Raptor
p70S6K
4EBP1
PIP3
P13K
IRS
INSR
LAR
INS
SHIP
SKIP
SOCS
PTP1B
JNK
IKK
Phosphatidyl inositol signaling system
+ps : serine phosphorylation
+py : tyrosine phosphorylation
-py : tyrosine dephosphorylation
MAPK signaling pathway
04910 3/31/09
(c) Kanehisa Laboratories

Appendix B

Concordance analysis

Table B.1: Result of concordance analysis examplified in six verbs. Corresponding frames/bioframes are presented in table 5.4. For ontological types, c is equal to *Substance* in Semantic Network and p is equal to process. Statistics for all the verbs investigated are available at *www.ruc.dk/~sz/Regrel/thesis.*

Verb	**Ontological types**	**Examples**
regulate	$c_1\, rels_{cp}\, p$ ~65% $c_1\, rel_{cc}\, c_2$ ~6% $p_1\, rel_{pp}\, p_2$ ~27% $p_1\, rel_{cc}\, c_1$ ~2%	(..)data demonstrate that pkc-mediated phosphorylation of p-gp *regulates* the activity of an endogenous chloride channel(...)
affect	$c_1\, rel_{cp}\, p$ ~63% $c_1\, rel_{cc}\, c_2$ ~10% $p_1\, rel_{pp}\, p_2$ ~27% $p_1\, rel_{cc}\, c_1$ ~0%	
reduce	$c_1\, rel_{cp}\, p$ ~39% $c_1\, rel_{cc}\, c_2$ ~9% $p_1\, rel_{pp}\, p_2$ ~40% $p_1\, rel_{cc}\, c_1$ ~13%	urapidil reduces blood pressure via blockade of peripheral... Pc: boiling greatly reduces the number of bacteria
remove	$c_1\, rel_{cp}\, p$ ~7% $c_1\, rel_{cc}\, c_2$ ~31% $p_1\, rel_{pp}\, p_2$ ~0% $p_1\, rel_{cc}\, c_1$ ~62%	
inhibit	$c_1\, rel_{cp}\, p$ ~80% $c_1\, rel_{cc}\, c_2$ ~10% $p_1\, rel_{pp}\, p_2$ ~10% $p_1\, rel_{cc}\, c_1$ ~0%	"...ethanol *inhibits* 3h-gaba release...", "...glp-1*inhibits* glucagon release...". "...lithium *inhibits* the enzyme glycogen synthase kinase-3..." "...rapamycin *inhibits* the kinase mTOR..." "...delta and mu opioid receptor activation *inhibits* spontaneous gaba release..."
stimulate	$c_1\, rel_{cp}\, p$ ~85% $c_1\, rel_{cc}\, c_2$ ~8% $p_1\, rel_{pp}\, p_2$ ~7% $p_1\, rel_{cc}\, c_1$ ~0%	

Appendix C

BioFrame descriptions

These following descriptions are frames and all roles are not nessesary for requiring the biomedical meaning of the word.

Remove(FN) *Agent/cause* causes *theme/(patient)* to move from initial location (*source*)covers "suicide inhibition". Patient of act is deleted from location and thus have no more activity here. It is perhaps mostly in biomedical texts this is true. *Example: Insulin/betacell secretion* causes *glucose transport* to happen from *blood* into the cells. (i.e. remove glucose from blood to cell)

Hindering(FN) *Hindrance* makes it difficult for *protagonist/(agent)* to do *Action.* covers inhibition. *Example: glp-1* makes it difficult for *body* to perform *glucagon release.*

Change_position_on_a_scale(FN) position of *item/patient* is affected on some scale. *Example:*The level of *insulin* is decreased.

Cause_change_of_position_on_a_scale(FN) *Agent/cause* affects position of *item* on some scale (*attribute*). In the molecular regulatory domain, the explanation can be translated to: *Agent/cause* affects position of *item/level* of patient on some scale (*attribute*). *Insulin/betacell secretion* causes *glycose* to lower the *blood sugar*.

Appendix D

Inhibition ontology in Description Logics

General Inhibition/Kinetics

$Kinetics \equiv \exists hasPart.Inhibition \sqcap \exists hasPart.Activation$

$Inhibition \equiv \exists partOf.Kinetics \sqcap \exists hasPolarity.Negative$

$AllostericInhibition \equiv Inhibition \sqcap \forall bindingSite.NotActiveSite$

$SubstrateInhibition \equiv Inhibition \sqcap \forall inhibitorOfProcess.Substrate$

$ProductInhibition \equiv Inhibition \sqcap \forall inhibitorOfProcess.ReactionProduct$

Irreversible Inhibition

$IrreversibleInhibition \equiv Inhibition \sqcap \forall restoresreactionrate.No$

$SuicideInhibition \sqsubseteq IrreversibleInhibition$

Reversible Inhibition (disjointness is missing)

$ReversibleInhibition \equiv Inhibition \sqcap \forall restoresreactionrate.Yes$

$InhibitionWithMichaelisConstantIncreased$
$\equiv ReversibleInhibition \sqcap \forall michaelisConstant.Increased$

$InhibitionWithMichaelisConstantUnchanged$
$\equiv ReversibleInhibition \sqcap \forall michaelisConstant.Unchanged$

$InhibitionWithMichaelisConstantDecreased$
$\equiv ReversibleInhibition \sqcap \forall michaelisConstant.Decreased$

$InhibitionWithMaxrateUnchanged$
$\equiv ReversibleInhibition \sqcap \forall maxrate.Unchanged$

$InhibitionWithMaxrateDecreased$
$\equiv ReversibleInhibition \sqcap \forall maxrate.Decreased$

$CompetitiveInhibition$
$\equiv InhibitionWithMichaelisConstantIncreased$
$\sqcap InhibitionWithMaxrateUnchanged$

$UncopetitiveInhibition$
$\equiv InhibitionWithMichaelisConstantDecreased$
$\sqcap InhibitionWithMaxrateDecreased$

$NoncompetitiveInhibition$
$\equiv InhibitionWithMaxrateUnchanged$
$\sqcap InhibitionWithMaxrateDecreased$

$MixedInhibition$
$\equiv InhibitionWithMichaelisConstantIncreased$
$\sqcap InhibitionWithMaxrateDecreased$

Disjointness:

$InhibitionWithMaxrateDecreased$
$\sqcup InhibitionWithMaxrateUnchanged$

$InhibitionWithMichaelisConstantIncreased$
$\sqcup InhibitionWithMichaelisConstantuUnchanged$
$\sqcup InhibitionWithMichaelisConstantDecreased$

Index

Bibliography

[1] Iso 704 2000 terminology work - principles and methods. international organization for standardization.

[2] OBO foundry. *www.obofoundry.org*, June 2010.

[3] K. Ahmad and L. Gillam. Automatic ontology extraction from unstructured texts. In Robert Meersman, Zahir Tari, Mohand-Said Hacid, John Mylopoulos, Barbara Pernici, Özalp Babaoglu, Hans-Arno Jacobsen, Joseph P. Loyall, Michael Kifer, and Stefano Spaccapietra, editors, *OTM Conferences (2)*, volume 3761 of *Lecture Notes in Computer Science*, pages 1330–1346. Springer, 2005.

[4] M. Ajspur and S. Zambach. Reduction of composites of relations between classes within formal ontologies. *ARCOE-11 Workshop Notes*, pages 26–30, July 2011.

[5] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, 28:381–390, Jul 2010.

[6] Sophia Ananiadou and John McNaught, editors. *Text Mining for Biology and Biomedicine*. Artech House, Boston, 2006.

[7] T. Andreasen and H. Bulskov. *On Deriving Data Summarization through Ontologies to Meet User Preferences*, pages 67–87. Springer Berlin / Heidelberg, 2009.

[8] T. Andreasen, H. Bulskov, P. A. Jensen, and T. Lassen. Conceptual indexing of text using ontologies and lexical resources. In *Proceedings of the Eighth International Conference on Flexible Query Answering Systems*. Springer, October 2009.

[9] T. Andreasen, H. Bulskov, T. Lassen, S. Zambach, P. Anker Jensen, B.l Nistrup Madsen, H. Erdman Thomsen, J. Fischer Nilsson, and B. Antoni Szymczak. SIABO - semantic information access through biomedical ontologies. In J. L. G. Dietz, editor, *KEOD*, pages 171–176. INSTICC Press, 2009.

[10] Troels Andreasen, Henrik Bulskov, Tine Lassen, Sine Zambach, Per Anker Jensen, Bodil Nistrup Madsen, Hanne Erdman Thomsen, Jørgen Fischer Nilsson, and Bartlomiej Antoni Szymczak. On semantic information access through biomedical ontologies. In *Proceedings of the Ninth International Conference on Flexible Query Answering Systems*, volume LNCS. Springer, October 2011.

[11] Aristotle. Metaphysics.

[12] A. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In *Proc. American Medical Informatics Assoc. (AMIA)*, pages 17–21, 2001.

[13] A. R. Aronson and F. M. Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17:229–236, May 2010.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.

[15] F. Baader, D. Calvanese, and D. McGuinness, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[16] F. Baader, C. Lutz, and B. Suntisrivaraporn. *CEL - a polynomial-time reasoner for life science ontologies*, pages 287–291. Springer Berlin/Heidelberg, 2006.

[17] F. Baader, C. Lutz, and A.Y. Turhan. Small is again beautiful in description logics. *KI - Künstliche Intelligenz*, 2010.

[18] R. Bach, Y. Iwasaki, and P. Friedland. Intelligent computational assistance for experiment design. *Nucleic Acids Res.*, 12:11–29, Jan 1984.

[19] Michael Bada and Lawrence Hunter. Desiderata for ontologies to be used in semantic annotation of biomedical documents. *Journal of Biomedical Informatics*, 44(1):94–101, 2011.

[20] Mark Balaguer. *The Stanford Encyclopedia of Philosophy, Platonism in Metaphysics*. ISSN 1095-5054. The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115, 2009.

[21] Elena Beisswanger, Vivian Lee, Jung-Jae Kim, Dietrich Rebholz-Schuhmann, Andrea Splendiani, Olivier Dameron, Stefan Schulz, and

Udo Hahn. Gene Regulation Ontology (GRO): design principles and use cases. *Stud Health Technol Inform*, 136:9–14, 2008.

[22] W. Blonde, V. Mironov, A. Venkatesan, E. Antezana, B. De Baets, and M. Kuiper. Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics*, 2011.

[23] Ward Blonde, Erick Antezana, Bernard De Baets, Vladimir Mironov, and Martin Kuiper. Metarel: an ontology to support the inferencing of semantic web relations within biomedical ontologies. In *International Conference on Biomedical Ontology, Conference proceedings*, pages 79–82. Nature preceedings, July 2009.

[24] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32:D267–270, Jan 2004.

[25] A. P. Bracken, N. Dietrich, D. Pasini, K. H. Hansen, and K. Helin. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.*, 20:1123–1136, May 2006.

[26] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, 29:365–371, Dec 2001.

[27] Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. Testing different ACE-style feature sets for the extraction of gene regulation relations from MEDLINE abstracts. In Tapio Salakoski, Dietrich Rebholz-Schuhmann, and Sampo Pyysalo, editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*, pages 21–28. Turku Centre for Computer Science (TUCS), 2008.

[28] Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. The genereg corpus for gene expression regulation events - an overview of the corpus and its in-domain and out-of-domain interoperability. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010.

[29] Bob Carpenter. *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge, England, 1992.

[30] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, 38:D473–479, Jan 2010.

[31] Ron Caspi, Tomer Altman, Joseph M Dale, Kate Dreher, Carol A Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G Shearer, Peifen Zhang, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*, 38(Database issue):D473–9, 2010.

[32] H. Chen and B. M. Sharp. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5:147, 2004.

[33] G A Churchill. Fundamentals of experimental design for cdna microarrays. *Nat Genet*, Dec 2002.

[34] Peter Clark, John Thompson, and Bruce W. Porter. *Knowledge Patterns*, pages 191–208. International Handbooks on Information Systems. Springer, 2004.

[35] The OBO consortium. Basic formal ontology (bfo). *ontology.buffalo.edu/bfo/*, 2010.

[36] Monica Crubézy, Mark A. Musen, Enrico Motta, and Wenjin Lu. Configuring online problem-solving resources with the internet reasoning service. *IEEE Intelligent Systems*, 18(2):34–42, 2003.

[37] D. A. Cruse. On the transitivity of the part-whole relation. *Journal of Linguistics*, 15:1–201, 1979.

[38] Ture Damhus, Peder Olesen Larsen, Bodil Nistrup Madsen, and Sine Zambach. Consistency and clarity in chemical concepts: How to achieve a codified chemical terminology - a pilot study. *Chemistry International*, 31(5):6–11, 2009.

[39] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation. *AI Magazine*, 14(1):17–33, 1993.

[40] C. Desler, S. Zambach, and M.FL. In silica characterization of hypothetical interactomes. *Manuscript*, page Will be submitted, 2009.

[41] A Dolbey, M Ellsworth, and J Scheffczyk. Bioframenet: A domain-specific framenet extension with links to biomedical ontologies. *Proceedings of KR-MED*, pages 87–94, November 2006.

[42] Andrew Eric Dolbey. *BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology.* PhD thesis, University of California, Berkeley, 2009.

[43] S. Eker, M. Knapp, K. Laderoute, P. Lincoln J. MESEGUER, and K. SONMEZ. Pathway logic: Symbolic analysis of biological signaling. *Pacific Symposium on Biocomputing 2002*, pages 400–412, 2002.

[44] Mariano Ferndndez, Asuncion Gomez-Perez, and Natalia Juristo. Methontology: From ontological art towards ontological engineering. Technical report, AAAI Technical Report SS-97-06, 1997.

[45] Charles J. Fillmore. The case for case. In Emons Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Wilson, New York, 1968.

[46] Charles J. Fillmore. Frame semantics and the nature of language. In S. Harnad, editor, *Origins and evolution of language and speech*, pages 155–202. Academy of Sciences, 1976.

[47] Luciano Floridi. *Information: A Very Short Introduction.* Very Short Introductions. Oxford University Press, New York, 2010.

[48] Gottlob Frege. *Begriffschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Halle:, 1879. Available in several translations.

[49] Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with DOLCE. In Asuncion Gomez-Perez and V. Richard Benjamins, editors, *EKAW*, volume 2473 of *Lecture Notes in Computer Science*, pages 166–181. Springer, 2002.

[50] Gene Expression Omnibus. Geo. geo and miame (minimum information about a microarray experiment). *http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html*, Feb 2010.

[51] Gene-Ontology-Consortium. GO ontology relations. *geneontology.org/GO.ontology-ext.relations.shtml*, May 2011.

[52] Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho, editors. *Ontological engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web.* Springer-Verlag, Heidelberg and Berlin, 2004.

[53] W3C OWL Working Group. OWL 2 web ontology language document overview. Technical report, W3C, October 2009. http://www.w3.org/TR/2009/REC-owl2-overview-20091027/.

[54] T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.

[55] Tom Gruber. Ontology. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1963–1965. Springer US, 2009.

[56] N. Guarino. Formal ontology and information systems. In N. Guarino, editor, *Formal Ontology in Information Systems (FOIS)*, pages 3–18. IOS Press, Amsterdam, 1998.

[57] N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In N. J. I. Mars, editor, *Towards Very Large Knowledge Bases*, pages 25–32. IOS Press, Amsterdam, 1995.

[58] Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5/6):625–640, 1995.

[59] J. A. Gulla, T Brasethvik, and H. Kaada. A flexible workbench for document analysis and text mining. *Lecture Notes in Computer Science*, 2004.

[60] Udo Hahn, Katrin Tomanek, Ekaterina Buyko, Jung-jae Kim, and Dietrich Rebholz-Schuhmann. 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*, pages 37–45, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[61] Petr Hajek. Fuzzy logic. In Edward N. Zalta, editor, *The Stanford Encyclopaedia of Philosophy*. Fall 2010 edition, 2010.

[62] K. Hansen, S. Zambach, and C. Theil Have. Ontology-based retrieval of bio-medical information based on microarray text corpora. In *Proceeding of ESSLLI 2010 Student Session*, Lecture Notes in Computer Science. Springer, 2010. Best Poster Award.

[63] K. R. Heidtke and S. Schulze-Kremer. BioSim–a new qualitative simulation environment for molecular biology. *Proc Int Conf Intell Syst Mol Biol*, 6(NIL):85–94, 1998.

[64] R. Heinrich and S.M. Rapoport. Metabolic regulation and mathematical models. *Prog. Biophys. Molec. Biol.*, 32, 1977.

[65] R. Hoffmann. Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds. *Curr Protoc Bioinformatics*, Chapter 1:Unit1.16, Dec 2007.

[66] Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, and Chris Wroe. *A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition 1.0*, August 2004.

[67] Cycopr Inc. The cyc ontology. *http://www.cyc.com/*, June 2008.

[68] Peter Ingwersen. *Information Retrieval Interaction.* Taylor Graham, London, 1992.

[69] Paul S. Jacobs, George R. Krupka, and Lisa F. Rau. Lexico-semantic pattern matching as a companion to parsing in text understanding. In *HLT*. Morgan Kaufmann, 1991.

[70] Thomas Jech. *The Stanford Encyclopedia of Philosophy, Nominalism in Metaphysics.* ISSN 1095-5054. The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115, 2002.

[71] Christopher Johnson, Miriam Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles Fillmore. *Framenet: Theory and practice.* Ebook, Berkeley, California, 2003.

[72] R. A. Cote K. A. Spackman, K. E. Campbell. Snomed rt - a reference terminology for health care. *Proc AMIA Annu Fall Symp*, pages 640–644, 1997.

[73] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[74] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, 35:D561–565, Jan 2007.

[75] A. Kilgarriff. Assorted frequency lists and related documentation for the british national corpus (bnc). *http://www.kilgarriff.co.uk/bnc-readme.html*, 1995.

[76] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182, 2003.

[77] Markus Krötzsch. Efficient inferencing for OWL EL. In Tomi Janhunen and Ilkka Niemelä, editors, *JELIA*, volume 6341 of *Lecture Notes in Computer Science*, pages 234–246. Springer, 2010.

[78] Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res*, 36(Database issue):D684–8, 2008.

[79] B.S. Laplace. *Translated from the fifth French edition of 1825.* New York: Springer-Verlag, 1995.

[80] T. Lassen and S. Zambach. Verb frequency lists, BNC, Medline abstracts and biomedical patents. *unpublished*, 2008.

[81] Bong-Soo Lee. Causal relations among stock returns, interest rates, real activity and inflation. *The Journal of Finance*, XLVII(4):1591–1602, 1992.

[82] C. Lindberg. The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc*, 61:40–42, May 1990.

[83] Brandon C. Look. *Gottfried Wilhelm Leibniz.* ISSN 1095-5054. 2008.

[84] C. Lutz and U. Sattler. The complexity of reasoning with boolean modal logic. *Proceeding of Advances in Modal Logic 2000 (AiML 2000)*, 2000.

[85] Carsten Lutz and Ulrike Sattler. Mary likes all cats. In Franz Baader and Ulrike Sattler, editors, *Description Logics*, volume 33 of *CEUR Workshop Proceedings*, pages 213–226. CEUR-WS.org, 2000.

[86] B.N. Madsen and H.E. Thomsen. *Terminological Principles used for Ontologies.* Litera, 2008.

[87] Bodil Nistrup Madsen and Hanne Erdman Thomsen. Terminological ontologies and normative terminology work. In *Proceedings of TSTT 2006, Third International Conference on Terminology Standardization and Technology Transfer*, 2006.

[88] Bodil Nistrup Madsen, Hanne Erdman Thomsen, Tine Lassen, and Sine Zambach. Insulinontologi til s_oe geprojekt. In *NORDTERM 2009: Ontologier og taksonomier*, volume 2, page 10, Copenhagen, Denmark, 2009.

[89] Hanne Erdman Thomsen Madsen, Bodil Nistrup and Carl Vikner. Principles of a system for terminological concept modeling. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 15–18, Vol. I 2004.

[90] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[91] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, 37:D619–622, Jan 2009.

[92] A. T. Mccray. An upper-level ontology for the biomedical domain. *Comp Funct Genom*, 9(4):80–4, 2003.

[93] Medline. National library of medicine. *www.ncbi.nlm.nih.gov/sites/entrez*, 2008.

[94] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[95] S. Moerk and S Zambach. Probabilistic ontologies. *Unpublished manuscript*, 2010.

[96] Hans-Michael Muller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 2004.

[97] Mark Musen. Bioportal. *bioportal.bioontology.org*.

[98] J. F. Nilsson, T. Andreasen, P. A. Jensen, P. Paggio, B. S. Pedersen, and H. E. Thomsen. Ontological extraction of content for text querying. In *Pre-proceedings of NLDB 2002*. Pre-proceedings of NLDB 2002, 7th. Int. Workshop on Application of Natural Language to Information Systems, June 2002.

[99] J. F. Nilsson, T. Andreasen, P. A. Jensen, P. Paggio, B. S. Pedersen, and H. E. Thomsen. Ontoquery: Ontology-based querying of texts. In *AAAI Spring Symposium*. AAAI Spring Symposium March 25-27, Standford University, California, March 2002.

[100] J. Fischer Nilsson. ONTOLOG - A Logico-Algebraic Framework for Ontologies. *Proceedings of the First International OntoQuery Workshop*, 1, 2001.

[101] J. Fischer Nilsson. Diagrammatic reasoning with classes and relationships. In *Manuscript*, 2011.

[102] J. Fischer Nilsson. Querying class-relationship logic in a metalogic framework. In *Proceedings of the Nineth International Conference on Flexible Query Answering Systems*, volume LNCS. Springer, October 2011.

[103] J. Fischer Nilsson, B. A. Szymczak, and P. Anker Jensen. Ontograbbing: Extracting information from texts using generative ontologies. In *Proceedings of the Eigth International Conference on Flexible Query Answering Systems*. Springer, October 2009.

[104] D. McGuinness O. Lassila. The role of frame-based representation on the semantic web. *Linkoping Electronic Articles in Computer and Information Science*, 2001.

[105] M. P. Oakes. *Concordancing, collocations and dictionaries*, pages 149–198. Edingburgh Textbooks in Empirical Linguistics. Edingburgh University Press, 1998.

[106] Seán I. O'Donoghue, Heiko Horn, Evangelos Pafilis, Sven Haag, Michael Kuhn 0004, Venkata P. Satagopam, Reinhard Schneider, and Lars Juhl Jensen. Reflect: A practical approach to web semantics. *J. Web Sem*, 8(2-3):182–189, 2010.

[107] C. S. Peirce. On the algebra of logic: A contribution to the philosophy of notation. *American journal of mathematics*, 7:180–202, 1885.

[108] Paul M. Pietroski. *Events and Semantic Architecture*. Oxford University Press, Oxford, 2005.

[109] Medline (Entrez Pubmed). Search data base for biomedical litterature. *http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed*, 2008.

[110] James Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441, dec 1991.

[111] James Pustejovsky, José M. Castaño, Jason Zhang, M. Kotecki, and B. Cochran. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing*, pages 362–373, 2002.

[112] Xiaoyan A Qu, Ranga C Gudivada, Anil G Jegga, Eric K Neumann, and Bruce J Aronow. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics*, 10 Suppl 5:S4, 2009.

[113] D. Ravichandran and E. H. Hovy. Learning surface text patterns for a question answering system. In *ACL*, pages 41–47, 2002.

[114] Gonzalo Rodriguez-Pereyra. *The Stanford Encyclopedia of Philosophy, Nominalism in Metaphysics.* ISSN 1095-5054. The Metaphysics Research Lab Center for the Study of Language and Information Stanford University Stanford, CA 94305-4115, 2007.

[115] D. L. Rubin, S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield, M. Ashburner, I. Sim, C. G. Chute, H. Solbrig, M. A. Storey, B. Smith, J. Day-Richter, N. F. Noy, and M. A. Musen. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, 10:185–198, 2006.

[116] Daniel L. Rubin, Nigam Shah, and Natalya Fridman Noy. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90, 2008.

[117] S. Schulz, M. Boeker, and H. Stenzhorn. How granularity issues concern biomedical ontology integration. *Stud Health Technol Inform*, 136(NIL):863–8, 2008.

[118] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, 25:1251–1255, Nov 2007.

[119] B. Smith, W. Ceusters, B. Klagges, J. Ko hler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biol.*, 6:R46, 2005.

[120] B. Smith and K. Munn. *Applied Ontology: An Introduction.* Ontos Verlag, Heusenstamm, 2009.

[121] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. *MEDINFO*, pages 444–448, 2004.

[122] Barry Smith and Thomas Bittner. Basic formal ontology for bioinformatics., April 01 2008.

[123] J. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[124] Paul Studtmann. *The Stanford Encyclopedia of Philosophy, Aristotle's Categories.* ISSN 1095-5054. The Metaphysics Research Lab Center for

the Study of Language and Information Stanford University Stanford, CA 94305-4115, 2007.

[125] SUMO. The Suggested Upper Merged Ontology. Teknowledge, 2003. Version 1.60.

[126] G. Sutcliffe and C. Benzmueller. Automated reasoning in higher-order logic using the TPTP THF infrastructure. *Journal of Formalized Reasoning*, 3(1):1–27, 2010.

[127] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou. Construction of an annotated corpus to support biomedical information extraction.

[128] R. Turner and A. Eden. The philosophy of computer science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2009 edition, 2009.

[129] P. L. Whetzel, H. Parkinson, H. C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S. A. Sansone, C. Taylor, J. White, and C. J. Stoeckert. The MGED Ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22:866–873, Apr 2006.

[130] W. A. Woods. *What's in a link? Foundations in Semantic Networks*, pages 32–82. Academic Press, New York, 1975.

[131] R. M. Woodsmall and D. A. Benson. Information resources at the National Center for Biotechnology Information. *Bull Med Libr Assoc*, 81:282–284, Jul 1993.

[132] C. T. Workman, H. C. Mak, S. McCuine, J. B. Tagne, M. Agarwal, O. Ozier, T. J. Begley, and T. Samson, L. D. andIdeker. A systems approach to mapping DNA damage response pathways. *Science*, 312:1054–1059, May 2006.

[133] S. Zambach. Towards ontology based search and knowledgesharing using domain ontologies. In *2nd Workshop on 3rd Generation Enterprise Resource Planning Systems*, volume 2, Copenhagen, Denmark, 2008.

[134] S. Zambach. A formal framework on the semantics of regulatory relations and their presence as verbs in biomedical texts. In *Proceedings of the Eighth International Conference on Flexible Query Answering Systems*, Lecture Notes in Computer Science, pages 443–452. Springer, October 2009.

[135] S. Zambach. Logical implications for regulatory relations represented by verbs in biomedical texts. In *International Conference on Biomedical Ontology, Conference proceedings*, pages 198–198. Nature preceedings, July 2009.

[136] S. Zambach and J. U. Hansen. Logical knowledge representation of regulatory relations in biomedical pathways. In Sami Khuri, Lenka Lhotsk, and Nadia Pisanti, editors, *Information Technology in Bio- and Medical Informatics, ITBAM 2010*, volume 6266 of *Lecture Notes in Computer Science*, pages 186–200. Springer Berlin / Heidelberg, 2010.

[137] S. Zambach, P. Holst, and Z. Worm Francker. Using corporate social responsibility strategy with a climate focus for enterprise systems. In Henrik Christiansen, editor, *RUC Sunrice Tripple C Conference*, Roskilde, Denmark, 2010.

[138] S. Zambach and T. Lassen. A lexical framework for semantic annotation of positive and negative regulation relations in biomedical pathways. In *Proceedings of the Fourth International Symposium for Semantic Mining in Biomedicine*, volume 714, pages 145–150, Cambridge, United Kingdom, October 2010. CEUR Workshop Proceedings.

[139] S. Zambach and B. N. Madsen. Applying terminological methods and description logic for creating and implementing an ontology on inhibition. In Jan L. G. Dietz, editor, *KEOD 2009 - Proceedings of the International Conference on Know ledge Engineering and Ontology Development, Funchal - Madeira, Portugal, October 6-8, 2009*, pages 452–455. INSTICC Press, 2009.