Procedurally Rhetorical Verb-Centric Frame Semantics as a Knowledge Representation for Argumentation Analysis of Biochemistry Articles

by

Mohammed Alliheedi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

External Examiner:	Vlado Keselj Professor, Faculty of Computer Science Dalhousie University
Supervisor(s):	Robert E. Mercer Professor, Dept. of Computer Science, The University of Western Ontario
	Robin Cohen Professor, School of Computer Science, University of Waterloo
Internal Member:	Jesse Hoey Associate Professor, School of Computer Science, University of Waterloo
Internal-External Member:	Randy Harris Professor, Dept. of of English Language and Literature, University of Waterloo
Other Member(s):	Charles Clarke Professor, School of Computer Science, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The central focus of this thesis is rhetorical moves in biochemistry articles. Kanoksilapatham has provided a descriptive theory of rhetorical moves that extends Swales' CARS model to the complete biochemistry article. The thesis begins the construction of a computational model of this descriptive theory. Attention is placed on the Methods section of the articles. We hypothesize that because authors' argumentation closely follows their experimental procedure, procedural verbs may be the guide to understanding the rhetorical moves. Our work proposes an extension to the normal (i.e., VerbNet) semantic roles especially tuned to this domain. A major contribution is a corpus of Method sections that have been marked up for rhetorical moves and semantic roles. The writing style of this genre tends to occasionally omit semantic roles, so another important contribution is a prototype ontology that provides experimental procedure knowledge for the biochemistry domain. Our computational model employs machine learning to build its models for the semantic roles and rhetorical moves, validated against a gold standard reflecting the annotation of these texts by human experts. We provide significant insights into how to derive these annotations, and as such have contributions as well to the general challenge of producing markups in the domain of biomedical science documents, where specialized knowledge is required.

Acknowledgements

This thesis would not have been possible foremost without the assistance and support of Allah (God) and then the help and support of many individuals in so many ways. At the top of the list of whom I wish to credit are my supervisors, Robert (Bob) E. Mercer, Robin Cohen, and former supervisor, Chrysanne DiMarco. Their knowledge and friendship were invaluable for the completion of this thesis. I would like to express my deepest appreciation to them for all of the things I have learned from them during the years I spent under their supervision in the PhD program. My supervisors have been always available and open to discuss and provide advice on different issues either on my research or outside the research area. I will never forget not only their positive attitude during the PhD years, but also their understanding and support when most needed. It was a genuine honor to work with all of them.

I also would like to express my gratitude and admiration to the members of my committee, Vlado Keselj, Jessy Hoey, Randy Harris, and Charles Clarke, for their guidance and encouragement throughout this process. I am especially grateful to the many people who contributed to this work in a variety of ways.

My appreciation and thanks extend also to my parents: Abdulrahman Alliheedi and Johra Alliheedi. My Mother (Johra) was the person who most encouraged and supported me to pursue a higher education. She was consistently available to offer a variety of support and advice. She will remain the greatest teacher in my life, from whom I learned a lot. I also cannot find enough words to express gratitude to my mother and my father for their presence in my life. I thank them for all of the support and patience they have shown since my first year in Canada to the present day. This thesis is dedicated to my parents especially my mother for her endless patience, her unconditional love, and encouragement. Her courage, support and confidence will always inspire me.

I am forever indebted to my family: to my wife, Suad, for her understanding, endless patience, and encouragement, support and confidence, and her unconditional love when it was most required; and to my children for being here at the right time, for their unconditional love, and for allowing me to have the time that I would have otherwise spent with them. I also thank Suad for being part of my daily life during my PhD years. Her presence has added valuable things to my life. Without her support and encouragement, completion of this thesis would not have been possible. She shared in all of the experiences I had in my PhD program day and night. I thank Suad for all of her support and encouragement. I am really blessed to have a wife like her. I wish also to express gratitude to my brother, Thamer, and sisters, Tahani and Nawal, for their usual supportive contacts and encouragement during my years in Waterloo. Their kind and uplifting words were a meaningful source of empowerment.

I would like also to thank all my friends for the great friendship we established during the PhD studies. We spent countless hours discussing different issues in research and academia in general, from which I benefited greatly. My thanks extend to all my colleagues for all of the enjoyable times we spent together.

Finally, I am also extremely grateful to Al Baha University and the Saudi Bureau for funding my PhD.

Dedication

This is dedicated to my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Scientists must routinely review the scholarly literature in their fields to keep abreast of current advances and to retrieve information relevant to their research. However, the volume of online scientific literature is immense, and rapidly increasing. In the biomedical field, the National Centre for Biotechnology Information (NCBI) developed a literature search engine, PubMed[1], to access various databases such as MEDLINE (journal citations and abstracts for biomedical literature), full-text life science e-journals, and online books. In 2010 PubMed repositories consisted of more than 20 million citations for biomedical literature [99]. By 2019 the number of citations had increased to more than 29 million[2]. As a consequence, it has become extremely challenging for biomedical scientists to keep current with information in their fields. This challenge has attracted Natural Language Processing (NLP) researchers to develop resources and automated tools for performing various tasks in Information Extraction (IE) and Text Mining (TM) using online corpora of biomedical articles, and thus enable biomedical researchers to better manage and exploit this volume of data [73]. These research activities have led to the development of a new field, Biomedical Natural Language Processing (BioNLP), a collaboration between the biomedical and computational linguistics/artificial intelligence communities [72].

The types of tasks currently handled by BioNLP systems have generally been aimed at extracting very specific and limited information, for example, protein and gene names and

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.ncbi.nlm.nih.gov/books/NBK3827/

relations [36], and so have been able to rely on relatively simple forms of information extraction. BioNLP has adapted various standard information extraction techniques, including both rule-based (e.g., shallow parsing, syntactic pattern-matching) and Machine Learning (e.g., Support Vector Machines, k-nearest neighbour classification method), to address several text-mining tasks, including extracting: protein-protein interactions (PPI) [87], drug-drug interactions (DDI) [140], gene relationships [74], and protein-residue associations [124].

But other, more in-depth and comprehensive, information contained in biomedical texts would be highly valuable to scientists because this type of information can enable validating scientific claims, tracing current research directions in their field, reproducing scientific procedures and so forth. Recently, a new and more challenging information extraction task has been introduced as a means of obtaining these types of detailed information: identifying the argumentation structure in biomedical articles (e.g., [62] and [63]). *Argumentation mining* can be used to validate scientific claims and experimental methodology, and to plot deeper chains of scientific reasoning. Unlike earlier simpler forms of information extraction, here the goal is to identify the structure of argumentative components within an entire text—for example, premises, evidence, conclusions—as well as the relationships between components.

1.2 The Problem Statement

Over the past decade, the focus on argumentation mining has been growing significantly in different areas of Artificial Intelligence (AI) research. The incentive to build Natural Language Processing (NLP) systems to automatically identify and analyze argumentative components in various genres of texts has increased because knowledge of argumentative structure facilitates various tasks such as text summarization [154] and opinion mining for commercial purposes [176].

The study of automated argumentation analysis has attracted the interest of several communities, including both scientific and computational linguistic researchers. Researchers from Argumentation Theory and Artificial Intelligence have come together to develop this new field of Computational Argumentation in interdisciplinary conferences and publications (e.g., Computational Models of Argument Workshop (COMMA), *Argument & Computation* Journal). Various computational studies have been done to analyze different argumentation aspects, including: the structure of valid arguments in legal documents [109], scientific articles [65, 62, 63], and the role of argumentation in multi-agent systems [112]. In particular, researchers are developing automated argumentation analysis

systems to enable scientists in the experimental sciences to review and evaluate scientific findings more efficiently, and to help identify whether scientific claims are valid or not, based on their argumentative structure [95, 151].

In addition to the biomedical field, researchers have worked on argumentation mining tasks in a variety of other domains, mainly: on-line debates [26], legal documents [109], newspaper articles and court cases [48], and product reviews [176]. However, these approaches have lacked consistency in their definitions of argumentation "schemes" (i.e., labels used to identify the different components of an argumentative structure). Moreover, there has been no formal, computationally feasible, semantics for these schemes. As a consequence, it has been difficult to build automated systems that can identify the components of an argument with a high degree of accuracy. And, because of the many different argumentation schemes, it has been impossible to come up with standardized metrics and evaluations of these different approaches.

In this research we will work on the biochemistry domain to develop a formal knowledge representation, *procedurally rhetorical verb-centric frame semantics*, that can be used for in-depth argumentation analysis, is computationally feasible to implement, and will enable argumentation mining of more-detailed scientific knowledge than is currently available. This will be an important step towards providing researchers in Computational Argumentation working in domains with similar discourse structure with a means of using and evaluating the metrics we will develop. To the best of our knowledge, no research has proposed or incorporated the idea of a semantic frame based on verb analysis to assist in the analysis of argumentation in biochemistry articles.

We will argue that verb-centric analysis is central to the understanding of biochemistry articles; this will explain why the knowledge representation that we choose to use is one that focuses on procedures. We will also clarify the importance of delineating the possible semantic roles associated with these verb frames, in order to produce an effective representation of the article's argument structure. In essence, a combination of rhetorical moves and semantic roles form the centrepiece of our proposed framework for natural language analysis.

We also explore the importance of developing an accompanying knowledge ontology as part of the semantic representation. Doing so enables effective inference of crucial domain knowledge that drives the understanding of the articles that are being processed.

A key element of our process of resolving the required components of rhetorical moves and semantic roles is the engagement of human annotators who are experts in the biochemistry domain. We in fact advocate for a critical partnership between possible end users of the natural language processing system (the experts) and NLP researchers who

have deep understanding of the challenges of the linguistic processing and what is can be supported as output and representation from these systems. As will be shown, it is in fact an iterative process of interaction between these two groups of people that is required in order to produce the needed grounding for the argument analysis of these texts. One central contribution of our work, therefore, will be to outline methods for effectively introducing human annotators into the process. In so doing, we also increase proper understanding of argument-related annotation schemes, with insights into mapping out the human-in-the-loop procedure, in general.

The effort that we devote to determining how best to engage with human annotators also crucially assists in enabling a validation of the knowledge representation that we propose for analyzing our biochemistry articles. Working with our experts, an annotated dataset is produced which can then be leveraged as the gold standard comparison when testing the effectiveness of our methods.

The structure of the document[3] will be as follows: First, an overview of some theoretical and computational approaches to argumentation are presented in Chapter 2. Chapters 3 through 6 present our proposed model in detail: the Rhetorical Moves, the Semantic Roles and Frames, the process of Annotation and the construction of the Ontology. Chapter 7 discusses the System as a Whole, as well. Finally, a discussion of lessons learned and future work alongside with a conclusion of this thesis is given in Chapter 8.

1.3 An Overview of Our Design Process

In the chapters that follow, we will provide a detailed introduction to the process of argument analysis and to the chosen domain of application, biochemistry articles. We will also clarify important technical terms such as semantic frame, semantic roles and rhetorical moves. Related work on these problems will be covered briefly as well, in order to give a sense of the current state of progress on the problem we aim to examine: how best to derive the argument structure of biochemistry articles.

Once our proposed model is described in detail, we will also reflect on the lessons learned during this design process and make clear how the approaches that we develop may be of value for researchers examining other, related problems (for example analyzing other scientific articles or improving the general process of argument analysis).

Before we reveal the details of our framework, it is useful to have a high level overview of the kind of iterative design process that was employed for this thesis, in order to make

[3]Our published papers point to earlier versions of some of the work presented here, namely [3, 4, 5, 6]

clear how the various components of the solution arose and were integrated.

This chronicle will also assist in drawing out some of our key components and their inherent value. We began with an interest in analyzing biochemistry articles and soon chose to narrow our attention to the Methods section of these documents, and to focus on the underlying rhetorical moves, desiring a formal knowledge representation.

We decided to make use of frame semantics, with the idea that extracting this and distinguishing this along with the rhetorical moves would enable a proper analysis of biochemistry texts.

We learned that there was a lot of expert knowledge to deal with. We connected with biochemistry experts who could enable us to learn about biochemistry in proper detail, though these individuals had no knowledge of linguistics. We ultimately concluded that what was most significant for each article were the underling verbs, which led us to advocate for a procedurally-oriented solution. But it also became clear that the semantic roles that we would be identifying needed to be specialized, due to this particular application area. In other words, a general VerbNet solution would not be sufficient for our particular setting.

We needed to properly understand the verbs that were occurring: what they were, how frequently they occurred, and such. And we learned that for a large number of instances, the verbs were ones that we in fact not in VerbNet.

We learned more about the typical verbs and their uses through interviews with our expert annotators. We learned for instance that certain subcategories of the Instrument class had specific usages. For example, there were catalysts for instrumenting a tool/enzymes. And there were measurements, of use when aiding digestion.

At this point, we had assembled 105 core articles to examine. In fact, there is a very important challenge for properly acquiring an effective corpus for any natural language processing task and in this thesis we elaborate briefly on how we arrived at our set of documents.

We wanted to get a handle on the underlying linguistic categories through an examination of examples with our expert annotators. Since these people were also going to be labelling sample texts by hand for us, to produce an annotated dataset used as our gold standard (when testing our computational models), we were at the point where we needed to make decisions about guidelines for annotators.

We assembled a valuable corpus of articles by looking at distinct journals and receiving subscriptions to access the articles, a total of 3500 articles in all. Certain verbs that came up again and again have particular associated semantic roles, which helped to inform our decisions about which knowledge representation to use.

With some initial guidelines, we allowed our annotators to do a test run – working with some training data. There were many iterations at this stage in order for the annotators to properly understand the computational linguistics task at hand; this in turn led to a revision of guidelines before the ultimate phase of launching annotators to produce the desired annotated dataset.

We were then at the point where we needed to create Verb Frames and their associated frame semantics. The process involved having expert annotators describe to us the verb in question and its intended usage – e.g the verb dilate and its definition and usage. We did this for about 39 different central verbs. We were looking at fairly huge documents as well.

In learning what the verb required, we were able to progressively determine categories such as condition, theme, location, instrument, or patient. The annotators provided the knowledge of the verb and we in turn enlightened them about the linguistics.

At this point an important observation arose: there was a gap. Some information that the experts had did not explicitly appear in the texts. This was because that knowledge was implicit/understood by them due to their expertise. This gap then suggested to us that what would be useful to have, for our knowledge representation, was an ontology. This ontology could then be used to infer knowledge or to inherit knowledge between the various levels.

In order to get a handle on how best to design the ontology, we explored a particular case study – the topic of gel purification. The ontology for this was built step by step. We saw how each step related to those immediately before and after or where there were instances of containment relationships. (This process was basically one of knowledge engineering).

In consultation with an expert on ontologies, we tried to ensure that the ones that we were building could be employed for general use. They were to be built with steps and states, so that any experimental procedure could be labelled as an instance of one of the items in the ontology.

We were always aiming to produce an annotated corpus, in order to make use of this as the gold standard for our computational analysis of biochemistry articles.

We decided that one important step was to resolve differences between annotators, so that we had a stable and valuable set of labels for our validation. It was a very time consuming process to address the differences between annotators, and to try to resolve this, adjusting the guidelines and knowledge representation. Once this was done however, the benefit was yielding a dataset that was quite valuable for validation/training/accuracy measurements.

At this point, we had already settled on representing rhetorical moves built upon a well-founded set of semantic roles, as features for the rhetorical moves model. There were various key patterns for the moves such as theme and instrument. There were various central labels for the domain such as Description of Method or Appeal to Authority. We were encouraged by literature such as the research of Teufel [155] which advocated for categories for every aspect of rhetorical moves. But this work had tried to map things out for all of experimental science; we knew that by drilling down to our domain of biochemistry articles, we could produce a more valuable framework, and one that could be tested more effectively as well.

We also ultimately decided that it would be valuable to perform independent evaluations of the components of our proposed framework: the semantic roles, the rhetorical moves, and the overall annotations. This would be done for an entire article.

At this point, we basically had completed both our design and our validation, had yielded some critical insights into how best to engage with annotators, and had produced a specific proposal for procedurally rhetorical verb-centric frame semantics, as a contribution for natural language researchers most interested either in rhetoric or the use of verb frames or the delineating of semantic roles. Our work suggests an important step towards processing scientific articles in general and also produces a valuable system for analyzing biochemistry articles in particular.

1.3.1 Support for Our Approach within the NLP Community

We end this chapter with a few key observations made recently by leading NLP expert Nancy Ide, revealed during her recent keynote address at the 2019 Canadian AI conference [76]. These viewpoints serve to reinforce the value of certain key design decisions that were taken in this thesis.

Several key NLP researchers to date (such as Nancy Ide keynote address at Canadian AI) [76] contend that there are unique challenges for scientific text mining, which suggest that the solutions being developed require a distinct set of design decisions. Issues that arise include difficulty with the terminology that exists in the texts due to heavy use of domain-specific words and phrases. Ide explains that it then is also difficult to obtain a specific gold standard corpus for validating the computational solutions that are developed. Even incorporating some kind of dictionary into the natural language processing is a challenge (time consuming, requiring domain-specific knowledge on the part of the designer). She also describes what she refers to as an "annotation bottleneck" [76]: enormous amounts of data are always desirable for the text processing task and this is even more pronounced

when examining biological articles. General purpose annotated corpora are simply not specific to the context of biology, requiring the development of specialized corpora. But gold standard corpora are then expensive to create, and this is also the case with supporting resources like ontologies.

We view our decision to devote considerable attention on developing a framework which describes how to effectively engage human annotators for our application of biochemistry articles, to be especially important. This assists in addressing the challenge of domain-specific terminology and in establishing the required ontologies for the natural language processing.

Ide also emphasizes the importance of drawing on expert knowledge in order to make the annotation process better. She advocates a vision of human-in-the-loop solutions and domain adaptation. In her view, it is important to do manual annotation for bootstrapping. This particular stance on the best path forward for NLP in the application of biological articles coincides well therefore with our specific strategy of engaging human experts.

Ide explains as well the importance of ensuring that the semantic categories employed in the computational solutions are well chosen. This viewpoint therefore also helps to confirm the value of our design decision to devote critical energy on the establishment of the set of semantic roles and rhetorical moves underlying the processing.

As we will explain in more detail in the final chapters of the thesis, while we are demonstrating our proposed approach in detail for the specific application of biochemistry articles, we do have general insights of use for any NLP researcher invested in making their repository of scientific articles more useful to end users, which we will draw out in more detail after presenting our proposed solution in full.

1.4 Contributions

There are two primary contributions arising from this thesis. The first is to provide important insights for the specific application area of biochemistry articles. We produce a detailed annotated dataset of articles, which clarify the central frames and their semantic roles as well as the rhetorical move argument structure. An ontology appropriate for this domain is also designed and created. The second is to deliver starting points for other computational linguistics researchers, especially ones who are invested in operating in biomedical domains. Our procedurally-oriented verb-centric frame semantics assists in mapping out this kind of knowledge representation and our ontology discussion draws out the importance of introducing additional domain knowledge. In addition, we have detailed

guidelines for annotators, derived from extensive hands-on experience, to shed light on how best to work with domain experts.

Chapter 2

Background

2.1 Argumentation Theory

In this section, we present a brief overview of argumentation, describe classical models of argumentation and discuss some of the early works on argumentation analysis.

2.1.1 What is Argumentation

Argumentation can be defined as "a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint" [166]. The essence of argumentation can be considered as influencing others to gain their adherence to a particular idea [120]. Tindale [159] defined argumentation as "the site of an activity, where reasons are given and appraised, where beliefs are recognized and justified, and where personal development is encouraged". Arguments have an explicit logical structure, for example, claims that are backed with reasons, which in turn are supported by evidence, leading to conclusions [160].

Argumentation analysis is the recognition and identification of the different forms of argumentative structures in texts. It is a crucial preliminary step for enabling the mining of in-depth argumentative elements in texts. This analysis enables, for example, a researcher to review, evaluate, or validate claims that are found in scientific articles. The difficulty in analyzing argumentation automatically is due to argumentative organization not being easily detected and recognized in texts, nor being well-determined (e.g., correlation with specific word types). Understanding argumentation requires deep analysis of texts to

identify its organization and logical structure. One type of knowledge that can be used to enable this deep analysis is lexical semantics.Various studies have used recurrent patterns of text organization called *moves* (i.e., text segments that are rhetorical and perform specific communicative goals) to analyze argumentative organization in texts manually [149], or automatically [154]. However, using these patterns with lexical semantic knowledge would provide additional information to more accurately detect and recognize the argumentative elements.

2.1.2 Classical Models of Argumentation

Argumentation has been studied throughout human history. In this section, we focus on the two most influential viewpoints: the classical view and the rhetorical view of argumentation. Perelman and Olbrechts-Tyteca believe that the goal of argumentation is to influence others to gain adherence of a particular idea or topic. Argument only exits when human minds communicate. Therefore, the most important factor of argumentation, as illustrated in "The new rhetoric: A treatise on argumentation" [120], is the audience. On the other hand, Toulmins study focuses on the understanding of the structures and the uses of argumentation in various fields such as philosophy, biology, law, and logic. Perelman and Olbrechts-Tyteca have spent some effort to understand the relationship between audiences and speakers. For instance, some scientific article authors do not need to worry about attracting an audience because their audiences are provided by the scientific institutions. This means that, one of responsibilities of scientists is to keep updated with recent scientific progressions. Thus, published articles will be read by other scientists regardless of how appealing the writing is to readers. However, in most cases, distribution techniques are not enough to guarantee readers, i.e., distributing magazines does not guarantee that people will purchase or read it. Therefore, to understand argumentation, Perelman and Olbrechts-Tyteca argue that we have to understand the listeners. The authors argue that the listeners may choose to listen to or not to listen to a speaker for various reasons. For example, a small child must listen to her or his parent while the parent may not want to listen to the child's reasoning. Perelman and Olbrechts-Tyteca have defined audiences as "the ensemble of those whom the speaker wishes to influence by his argumentation." [120]. Furthermore, various factors define the characteristics of audiences. Such factors include psychological background, social background, social functions exercised by the audience. For this reason, speakers may choose to divide her or his audience into groups based on social background, religious differences, etc. If the speaker holds a strong belief in her or his argument, however, the the audience may be viewed as a single entity, because the speaker believes everyone should be convinced of her or his opinion on this matter. This leads to the concept of universal audience. Perelman and Olbrechts-Tyteca have defined

universal audience as follows: "such as audience consists of the whole mankind, or at least, of all normal, adult persons" [120]. To a speaker, universal audience can be the conceptual audiences that they must consider and appeal to regardless of their targeted audiences. The concept of universal audience can also be used to test the absolute validity and the self-evident nature of an argument. Moreover, one may argue that an argument which targets the universal audience is objective while an argument that targets a particular group of audience is subjective. A universal audience is both general and conceptual. Universal audience exists in speakers' minds to help speakers refine their argument. On the other hand, the audience is more concrete. Speakers can identify a group of people in the society as a type of audience while universal audience only exists conceptually. It is difficult to convince universal audience rather than targeted audiences. Toulmin has not focused on the relationship between the audiences and argumentation [160]. Toulmin focuses on the structures of argumentation itself. To this end, Toulmin has identified the most fundamental elements of an argument as datum (D), warrant (W), and conclusion (C). Toulmin believes that the most common form of an argument is "Given D; since W; C" [160]. The warrant explains why we can draw the conclusion from the datum. Toulmin further expands this basic form with qualifier (Q) and rebuttal (R). A qualifier can be used to represent the uncertainty of a conclusion based on the given datum and warrant, while rebuttal can indicate circumstances that a warrant cannot be applied. Furthermore, backing (B), which is introduced as the last component of an argument, is the evidence of a warrant. Toulmin's model mainly aims to gain acceptance without requiring the truth. The Aristotle model, however, relies on truthfulness for one to build her or his arguments to appeal to others [130]. Comparing to the Perelman and Olbrechts-Tyteca model, we found that the Toulmin model is more mechanical and practical since this model provides us a framework for creating and identifying argumentative structures in everyday life. On the contrary, we found that the Perelman and Olbrechts-Tyteca model relies heavily on the notion of the universal audience, which is fundamentally abstract and vague since there is no clear definition of the universal audience that people can consider when they build their own arguments.

2.1.3 Rhetorical Approaches to Argumentation

Swales [149] proposed the Create-A-Research-Space (CARS) model that uses intuition about the argumentative structure of scientific research articles. Swales defined rhetorical moves as text segments that convey communicative goals. He reviewed the Introduction section in 48 articles from social and natural science and found common rhetorical structures among most of these articles. Swales identified three moves in these articles: establishing a research territory, establishing a niche, and occupying the niche. However,

despite the widespread influence of the CARS model, some researchers observed two problems: (i) the inconsistent assignment of rhetorical moves to text segments because the identification of the rhetorical moves relies on overall text comprehension, and (ii) a lack of empirical validation of moves in linguistic terms [79].

To overcome these problems, Kanoksilapatham [79] advanced Swales' approach to move analysis by developing a framework that combines his original CARS model with the use of Biber's multidimensional analysis [14] to enrich the model with additional information about linguistic characteristics. Biber's multidimensional analysis [14] is concerned with variation in the speaking and writing of English. Multidimensional analysis can be used to identify differences in linguistic characteristics between various text types at different levels of document structure (e.g., genre, internal section level). Although Kanoksilapatham provides an extension to the Swales move analysis study, and attempted validation of these moves in biochemistry articles, she only provides a descriptive analysis about rhetorical moves without defining an explicit method for analyzing and recognizing these moves in texts.

Gladkova [57, 58] did a detailed study to identify features that can be linked to argumentative organization in texts. Gladkova's argumentation structures, *topoi*, draw on classical argumentation theory [130]. Gladkova's findings show that argumentative organization is not correlated just by isolated linguistic features but rather with their stylistic *configurations*. The elements of these configurations included lexico-grammatical and semantic relations, syntax, deixis, and coreference. There is a key difference between these well-defined stylistic configurations and the usual loose collections of stylistic features in Machine Learning NLP. Gladkova's features of stylistic configurations interact with one another and with their semantic and syntagmatic environments in rich but regular ways [59]. Although Gladkova's corpus was not annotated by linguists other than herself or by domain experts, since the corpus was small, it would be feasible to include guidelines on how to annotate topoi, as suggested by Cohen et al. [35]

Walton et al. [170] developed a list of argumentation schemes for argumentation analysis. These schemes, forms of argument, aimed to represent common types of arguments including indicative, deductive, and abductive (defasible) arguments. Walton et al. defined a defeasible argument as "A defeasible argument is one in which the conclusion can be accepted tentatively in relation to the evidence known so far in a case, but may need to be retracted as new evidence comes in." [170]. The purpose of argumentation schemes is to study and analyze defeasible arguments in everyday life. An argumentation scheme is defined as the form of premises, conclusion, and related critical questions in an argument type. The notion of argumentation schemes is not relatively new. On the one hand, Hastings [66], systematically analyzed schemes for defeasible argumentation in his PhD thesis.

Kienpointner [81], on the other hand, developed a comprehensive listing of the schemes for deductive, inductive, and defeasible argumentation. Furthermore, Walton [171] identified 26 argumentation schemes for defensible argumentation and attached related critical questions to each of these schemes. These critical questions are the key to analyze, validate, or disapprove a given defeasible argument. Walton et al. [170] provided concrete examples of different defeasible arguments associated with their critical questions. As demonstrated in these examples, critical questions enable us to analyze how strong or weak an argument is. The answers to critical questions can be used to judge if an argument is valid or fallacious. The authors have argued that enthymemes may be historically misinterpreted and enthymemes could have meant defeasible (presumptive) argument schemes in the original Aristotelian meaning. Furthermore, argumentation schemes are important in pedagogy. Premises, conclusion, and critical questions give interlocutors, analysts, teachers, and students ways to study and understand argumentation theory. One of the most popular approaches is to draw diagrams, which display the relation between premises, conclusions and critical questions. The practice of drawing diagrams to understand arguments has been recognized in court and school. The authors further explained possible problems of argumentation schemes. Critical questions in argumentation schemes may have a completeness problem. That is, critical questions can be asked indefinitely. The authors have argued that argumentation schemes are important in the field of artificial intelligence and the first step towards formal understanding of argumentation is to precisely define argumentation schemes. Although these schemes aimed to represent common types of arguments including indicative, deductive, and defeasible arguments, these schemes were not intended for scientific arguments.

Overall, these different approaches based on argumentation theories for analyzing and recognizing argumentative elements, including move analysis [79, 149], argumentative zoning [153], and epistemic topoi [57], lacked a formal knowledge representation which could be used computationally for in-depth argumentation analysis and mining. Another problem in identifying argumentative elements is that relatively few biomedical related corpora annotated with argumentation structures currently exist for use in training or evaluating Machine Learning classifiers.[1] This has encouraged researchers to begin developing annotated corpora for use by the Computational Argumentation community ([62, 63], in particular). In the next section, we will give an overview on some of the state-of-art approaches in computational argumentation including annotation schemes for argumentative texts, extraction of argumentative structures in legal documents, detection of argumentative relations in debate corpus, and others.

[1]We note, however, increasing attention to this concern, with the design of such corpora as The Internet Argument Corpus (IAC) for research in political debates on internet forums [168] and the Dr. Inventor Multi-Layer Scientific Corpus (DRI) for computer graphics articles [91].

2.2 Computational Argumentation

In this section, we describe some of the state-of-the-art approaches in computational argumentation in two main themes: recognizing schemes and detecting argumentation.

2.2.1 Approaches for Recognizing Argumentation Schemes

Argumentative Zoning (AZ) was developed by Teufel and Moens [153] to categorize sentences based on their contextual information (e.g., determining authorship of knowledge claims). The AZ scheme classifies sentences into seven categories including the ones from the CARS model [149]. The data set consisted of 48 computational linguistic papers. Three annotators were involved in the study to extract sentences that fell into these seven categories. The results showed a Kappa score of 83% and 82% between the annotators in the first and second schemes, respectively. The AZ scheme was later modified to suit the characteristics of biology articles [108]. Furthermore, Teufel [151] proposed a revised version of AZ to include more new categorizes for annotating scientific articles such as chemistry. This revised version was planned to model all experimental sciences, which is challenging, since the style of scientific writing varies across disciplines.

Feng and Hirst [48] proposed an approach for recognizing argumentation schemes in the Araucaria corpus [132] that consisted of over 600 manually annotated arguments with their internal structures, premises, and conclusions. These arguments were from various sources including newspapers and court cases. Using the internal structures of arguments identified by the human annotators, the authors developed a method for recognizing the schemes in these arguments and classifying them into their proper categories accordingly. The authors used a set of common argumentation schemes described in [170] which include: argument from example, argument from cause to effect, practical reasoning, argument from consequences, and argument from verbal classification. The authors used statistical classifiers (i.e., one-against-others and pairwise) to classify the arguments into their appropriate schemes. Although, the system achieved accuracies slightly over 90% in classifying annotated arguments in only two of the argumentation schemes, argument from example and practical reasoning, the system performed poorly in classifying other schemes such as argument from consequences and argument from verbal classification.

Liakata et al. [95] developed an annotation scheme called Core Scientific Concepts (CoreSC) to classify sentences into scientific categories (e.g., related to authors other work). The CoreSC scheme consists of three layers: the first includes several categories to classify sentences; the second layer is concerned with properties of these categories; and the third layer creates a link to related instances of the same category. The authors use Machine

Learning classifiers (i.e., Conditional Random Fields and Support Vector Machines) to automatically classify sentences into the CoreSC categories. The data set consisted of 265 biochemistry and chemistry articles. The authors were only able to achieve an accuracy around 50% in categorizing sentences in the appropriate CoreSC scientific categories which is inadequate for such a task.

Green [62] proposed a plan for creating an annotated corpus of biomedical genetics research articles. Green emphasized that this corpus would be beneficial to the argumentation mining community since it would provide a fine-grained annotation of argumentative components. Also since there are as yet few annotated corpora available, such a corpus would enrich research in the field of Computational Argumentation in general. The author stated that this corpus will be publicly available for further investigation by different research groups in various tasks of argumentation mining.

Green [63] specified a set of argumentation schemes for scientific claims in genetics research articles. The author used a corpus of unannotated genetics research articles, and identified the components (e.g., premises, conclusions) of an argument as well as its type of scheme. Based on the analyses of various genetics research articles, the author specified 10 argumentation schemes that are semantically different. These schemes were new and had not previously been proposed. Furthermore, the specification of argumentation schemes was used to create annotation guidelines. Then, these guidelines were evaluated in a pilot study based on participants' ability to recognize these schemes by reading the guidelines. Overall, the author's ultimate goal for this initial study was to develop annotation guidelines for creating corpora for argumentation mining research. However, based on the pilot study, the results showed a variation in performance since there were two groups of participants (i.e., undergraduate students and researchers). The students performed poorly in recognizing argumentation schemes while the researchers were able to identify these schemes correctly in most cases.

Kirschner et al. [85] proposed an annotation scheme to identify argumentative structures on a fine grained level in scientific articles. This scheme includes four types of binary argumentative relations between sentences. These relations include consist of directed relations (e.g. attack, support, and detail) and an undirected relation (e.g., sequence). Moreover, the authors created a corpus to include 24 articles from educational psychology developmental psychology domains. The authors also developed a web-based annotation tool called DiGAT to support the annotation of argumentative components and their relations, and to visualize the argumentative structures as graphs. Four annotators were involved in annotating the argumentative structures in the corpus using the annotation schemes. The authors developed a graph-based agreement measure to calculate the inter-annotator agreement. As a result, the authors evaluated the annotated corpus using this

agreement measure and found that the inter-agreement scores between the annotators are fair to low. Thus, the authors provide a qualitative study to investigate the reasoning behind those low agreements between annotators. Finally, the authors stated that the ambiguity of argumentation structures is the main reason for the low agreement scores between the annotators.

2.2.2 Approaches for Detecting Argumentation

Legal documents

Mochales and Moens [109] proposed a multi-layer approach to detect argumentation in legal texts. These layers included the detection of argumentative information, argument boundaries, relationships between arguments, and the classification of argumentative elements, either as a premise or conclusion. The data set is comprised of legal documents from the European Court of Human Rights (ECHR) corpus. The authors achieved an accuracy of 80% in detecting argumentative units. They also achieved scores between 68% and 74% F1 on the classification of premises and conclusions, respectively. Finally, the last layer detected the argumentation structure by manually parsing the texts using context-free grammar (CFG) rules, achieving an accuracy of 60%.

Newspaper articles

Kiesel et al. [82] proposed a shared task to manually identify argumentative structures in newspaper editorials. The authors proposed a dialectical model of argumentation to identify explicit argumentative units (e.g., a claim and premise) and implicit argumentative relations (e.g., attack or support). The authors also created a corpus of newspaper editorials from three sources: Al Jazeera, Fox News, and The Guardian. Then, the corpus is annotated by identifying the topics, the argumentative units, and the argumentative relations. Finally, the authors proposed an evaluation measures to be used for further extension of the current study.

Lawrence and Reed [92] proposed to aggregate three different methods of extracting argumentative structures from texts. These methods are indicators of discourse, topical similarity, and a supervised machine learning approach based on argumentation schemes. The discourse indicators are used to identify argumentative relations between consecutive propositions. Thus, the authors only looked at specific terms to convey different relation types (e.g., support, conflict). For example, words like because, therefore, and since are associated with a support relation, but words like however, though, and nonetheless are

indicators of a conflict relation. In addition, the authors proposed a topical similarity which is similar to the one proposed in [93]. The topical similarity is based on the idea that every argument structure can be represented as a tree, and this tree is generated depth first [92]. So, an argument conclusion is stated first and followed by a line of supportive reasoning. Once the line of supportive reasoning is made, the argument moves in the tree to support another point. Based on this view of argument structures, the authors argued that determining an argument structure can be done by looking to the topical similarity between a proposition and its predecessor. If both are similar, then they are connected and vice versa. The authors employed WordNet to identify the similarity between a set of synonyms of each word in both propositions. Furthermore, the authors used a Naïve Bayes classifier to automatically identify components of two types of Walton's argumentation schemes (e.g., Expert Opinion and Positive Consequences) in unanalyzed texts. Then, the authors used these components to identify the presence of a specific scheme. Overall, the authors demonstrated that combining different automatic techniques of argumentation mining can accomplish higher results which can be closely compared to the manual analysis of a particular text.

Debate discourses

Cabrio and Villata [26] proposed using a textual entailment approach to detect and identify relationships between arguments in debate discourse. The corpus used in their study was on-line dialogues from Debatepedia, an online resource of arguments on critical issues. Textual entailment infers a directional relation between two text parts. The concept underlying textual entailment is the identification of the correlation, either support or contradiction, between two text segments. For a pair of text segments to be related by entailment, there must be a relation between the segments, termed "Text and Hypothesis", where the initial segment ("Text") is the first part of the argument (entailment) and the second segment ("Hypothesis") is the second part of the argument that either supports or contradicts the first part. In Cabrio and Villata's work, there was no manual identification of the entailment relationships between arguments. However, the authors used Dagan et al.'s [38] approach to defining and detecting textual entailment to infer these relationships. Then the authors identified the accepted arguments using Dung's argumentation theory framework [45]. In this framework, an argument is accepted when all arguments attacking it are rejected. However, an argument would be rejected if one of the attacking arguments is accepted. The result showed an accuracy of 75% in assigning a relation to a pair of arguments which reflects the total number of accepted arguments. However, the data set was too small and included only 200 T-H pairs (i.e., 100 T-H pairs were used to train the system and 100 T-H pairs to test it).

Bilu et al. [15] proposed an algorithm that automatically generates a negation statement based on an input claim. These statements can help to determine the validity or plausibility of the original claim. The proposed algorithm uses POS taggers to identify the main verb in a claim and adds negation words such as not into the claim. The algorithm further determines the usability of the negated claim using a logistic regression classifier with features such as word count and POS tags from the original claim. Through evaluation, it is shown that this algorithm can achieve as high as an 80% accuracy in generating clear and grammatically-correct negations. However, the algorithm itself is naive and there is still space to improve the sophistication of the negated claims.

Yansae et al. [177] also worked on the auto generation of debate arguments. However, their focus is to automatically determine the sentence ordering for a set of related debate argument segments. They identified that most claims are followed by one or more support sentences in a debate argument, stating that a clear argument should have a leading claim followed by support sentences. Therefore, learning this claim-support sentence structure is useful in automatically generating debate arguments. They formalize this ordering problem as follows. The input is a set of sentences containing a claim and one or more support sentences. The problem is to identify the claim statement and order the support sentences in a meaningful way. They formulated the identification of the claim as a binary-classification problem and modelled the ordering problem as a ranking problem. Both of these problems are solved using classic machine learning algorithms. In their evaluation, they reported the accuracy of the claim identification is only about 40%, leaving space for improvement. However, the authors believe identification of the claim statement in a debate corpus is an important step towards identifying prominent arguments in debates.

Boltuzic et al. [21] also contributes to argumentation techniques in online debates. The goal of their proposal is to identify prominent arguments in online debates. However, the authors argued that the first step is to classify online debate arguments with respect to their topics. Therefore, they proposed to combine clustering techniques with semantic textual similarity [1], which captures the degree of semantic equivalence between any two pieces of text. In particular, the inputs to this algorithm are the debate arguments extracted from online resources. A semantic textual similarity system is then used to determine similarity scores between pairs of arguments. These pair-wise scores are fed to the clustering algorithm and provide the final results of argument clusters sorted by their topics. Their evaluation has shown that in the best v-measure, 0 being worst case and 1 being perfect cluster, is only 0.3. This work has identified potentially challenges in automatically analyzing online debate arguments. In particular, we need better clustering algorithms to identify argument topics, as this is only the first step.

Reisert et al. [126] proposed a computational model to generate arguments from the

texts using the Toulmin model. The goal of automatic argumentation generation is to generate coherent and logically structured arguments. For every claim, there are statements that either support or attack that claim. Thus, the authors attempted to identify these supportive statements of both claims and counterclaims. Therefore, the authors created a set of rules to generate argumentative elements, which include data and warrant, and constructed a knowledge base of causality relations such as promote and suppress to represent negative or positive sentiments in these elements to a given claim. For example, supposing the claim is This House believes alcohol should be banned, so the causality relation, here, would be suppress (House, alcohol). A relation like promote (alcohol, liver disease) would be extracted from a supportive statement, or data, such as Alcohol causes liver disease. Furthermore, the link between these two relations would be the warrant in statement like this If the alcohol causes liver disease, then the House should ban it. Thus, the relation for that statement would be If promote (alcohol, liver disease), then suppress (House, alcohol). In conclusion, the results showed that the proposed system of generating arguments did not perform well and the system required substantial improvements to properly extract coherent and logically structured arguments.

Oraby et al. [114] proposed to investigate the characteristics of various styles for emotional and factual argumentation in online debates. The authors used a corpus of forum posts, namely, the Internet Argument Corpus (IAC) which includes manually annotated textual pairs of quote-response [168]. For every pair, the response was annotated either as a factual or emotional argument. So, the authors study the differences in argumentation styles in this corpus to extract linguistic patterns that are highly correlated with factual and emotional argumentation. The authors were able to identify these patterns using a supervised learner system called AutoSlog-TS [128]. Thus, the authors apply these patterns to a larger corpus of unannotated forum posts to obtain new linguistic patterns. Overall, the authors were able to derive different syntactic forms that are associated with factual and emotional arguments to classify various styles of them.

Wyner et al. [175] proposed an interactive tool to reconstruct and extract arguments called Argument Workbench. This tool was designed to be used by an expert in argumentation modelling called argumentation engineer. The Argument Workbench provides the engineer with different automatic processes to capture the information needed to facilitate the extraction of arguments. These processes include harvesting and pre-processing comments; highlighting argument indicators, speech act terminology, epistemic terminology; modelling topics; and identifying domain terminology and relationships. In this study, a corpus of texts discussing the Scottish Independence vote in 2014 is used. Furthermore, the authors used a conceptual semantic search over the corpus to extract sentences that are related to and an argument and domain terminology. With the extracted informa-

tion from the Argument Workbench, the argument engineer analyses this information and then inputs the analysis outputs into visualization tool called DebateGraph to layout arguments. Overall, the Argument Workbench tool is able to automatically capture the related information of a topic to facilitate constructing arguments manually.

Argumentation mining for the social web

Park and Cardie [116] developed a framework to automatically identify and classify argumentative components in online user comments. The authors identified four main categories (e.g., unverifiable, verifiable, verifiable public, and verifiable private) to classify a proposition, which is an elementary unit of argumentation. The authors manually annotate propositions in a corpus including over 1000 user comments extracted from an eRulemaking platform. The authors achieved inter-agreement of unweighted Cohens k score of 73% in only one third of the corpus. Furthermore, the authors employed a Support Vector Machine classifier to automatically classify each proposition into the four categories. The results showed that the automatic classification achieved a macro-averaged F1 score over 68%. Overall, this framework showed a promising result which will allow further investigation by integrating other aspects such as the identification of proposition relations in an argument.

Ghosh et al. [56] were among first group of researchers who attempted to tackle the argumentation mining problem in online interactions, including online blogs, online forums, and webpage comments. They determined that the first feasible step is to create an annotated corpus. They proposed to build this corpus through a multi-step process with human annotator. The first step involves expert annotators providing coarse-grained analysis. That is to say, the annotators will identify segments of texts that form a claim-attack-statement or claim-support-statement relations. The second step, they proposed using crowdsourcing and novice annotator to perform fine-grained analysis. For instance, a novice annotator is responsible for determining if a given relation is the attack or the support type. Through experimental evaluation, they have shown that their proposed method is scalable and achieved a reasonable accuracy with a F1 score of 62.6%.

Peldszus and Stede [119] proposed solutions to detect counter-considerations in texts. Counterconsiderations are pieces of text that authors can use to pre-empt potential counter-arguments. They argue that the detection of these counter-considerations is important because they are common in argumentative texts. Furthermore, these counter-considerations can be viewed as a special form of the support statements. They modelled this problem as identifying the role of the author in a piece of text. In particular, the author can either take a proponent role, with the regular support/claim statements, or take an opponent

role, with the counter-consideration statements. They used gradient descent learning to train a linear log-loss model for classification purposes. Through evaluation, they have shown that their model achieves an F1 score of 66% for newspaper articles. We believe that the authors have picked an interesting topic to explore, as counter-considerations are a special type of support statements. This work reveals the need to classify the types of support statements.

Sardianos et al. [134] have studied argumentation mining for unstructured online text. In particular, they propose a new algorithm to extract premises and claims from online news and online blogs. The authors claim that because the social web contains a huge amount of data, existing sentiment analysis can only determine the sentiment polarity for a particular topic. However, these analyses cannot capture public opinions, which are crucial for government policy decisions. To solve these problems, one feasible first step is to extract premises and claims from online corpus. Therefore, the authors have proposed using Conditional Random Fields to classify text segments. To keep the algorithm general, the authors only used PoS tags and language cues as training features. The F1 score of the results is only 32%. Despite this, we believe this is an important first step toward argument extraction in general online corpus.

Sobhani et al. [143] also contributed to the argument extraction from online news. They identified one of the major problems in argumentation mining is the lack of annotated corpus. They proposed a new framework, Non-Negative Metric Factorization, which requires only minimum learning supervision. In particular, their framework will map online corpus to various topics with minimum annotated arguments. Through evaluation, they have shown that their framework achieves an F1 score as high as 51%. The clustered argument can be further analyzed automatically to determine the public stance in a particular topic. We believe this clustering technique can be combined with other techniques to achieve better results.

Carstens and Tonis [29] proposed to extract arguments by identifying the attack or support relations between segments of texts. They argued that most argumentation mining techniques treat argument extraction as a two-step process. First, we must identify the argumentative text segments. Second, we analyze the relations between the identified segments. In their work, they proposed to combine these two steps into a single task because attack or support relations between text segments can be treated as evidence that a particular text segment is argumentative. Their work is backed up by the theory of argumentation framework proposed by Dung [45]. Dung's abstract argumentation theory captures the attack relations between the text segments. Carstens and Tonis were building an annotated text corpus to evaluate their proposed methods. This work is only at its initial stage and we are looking forward to future results.

Park et al. [117] targeted arguments in online user comments. They identified that most online user comments are argumentative claims without support. For instance, a user may state that the air travel fees should be more transparent without any support. To this end, as the first step in determining the validity of a claim, Park et al. propose to identify the types of supports required for any given claim. In particular, they have defined three types of supports: unverifiable, verifiable non-experimental and verifiable experimental. The unverifiable identifies the claims that cannot be presently verified. For instance, one cannot verify the outcome of a future event. Claims that only require the support of objective evidence are the verifiable non-experimental type. On the other hand, claims that must be supported by subjective evidence such as expert testimony are the verifiable experimental type. In 2014, Park and Cardie [116] proposed to use the support vector machine technique (SVM) as a classifier to determine the types of support for claims. This work has achieved an F1 score of 69%. Park et al. continued this work in 2015 and proposed to use Conditional Random Fields (CRF) [88] to exploit the sequential nature of the online user comments. However, the overall performance of CRF is not as good as SVM due to the heavy skew of claim types in the dataset. The authors think this support type identification is similar to argument scheme selection in a more general sense.

Arguments in persuasive essays

Song et al. [146] explored how to systematically facilitate annotation of arguments in persuasive texts. They developed an annotation protocol for annotators to classify arguments and their schemes in argumentative essays. They included an argument analysis task in their graduate admission tests. A student taking the test is required to critically evaluate arguments by annotating the given arguments. The annotations are then evaluated to match against argument schemes, which are proposed by [171]. They found that annotation is a laborious and it requires substantial training to achieve reasonable skill levels. This study has shown that argumentation schemes alone are not enough to provide the foundations in successfully annotating large corpus. Annotators or analyzers must be trained to achieve good performance.

Nguyen and Litman [111] proposed an approach to identify arguments in persuasive essays. This approach is based on the idea that separating argument words (e.g., view, conclude, and think) from domain words (e.g., art, and life) using LDA algorithm [18]. A corpus of 90 annotated persuasive essays is selected from eassyforum.com. Sentences in the corpus are annotated for three types of argumentative components: Major claim is the main author claim, claim, which is a statement that either supports or attacks a major claim, and premise which supports the validity of a claim. Furthermore, the authors

compared the proposed system against a baseline system [148] . In conclusion, the results showed that the proposed approach outperform the baseline.

Arguments in product reviews

Wyner et al. [176] proposed a tool that highlights potential argumentative sections of a text. The authors developed an argumentative analysis tool that consists of 5 tiers, including: the Consumer argumentation scheme, Discourse indicators, Sentiment terminology, the User model, and the Camera domain. The consumer argumentation scheme is used to represent the arguments that are related to "a course of action relative to preferences and values"[176]. Two factors were involved, the domain model and the user model, in the consumer argumentation scheme. The discourse indicators are used to identify linguistic expressions of discourse that show the relations between sentences. For simplicity, in this study, the authors considered only explicit indicators. Three types of discourse indicators were used: Indicators of premise, which include words such as "after", "as", and "because"; indicators of conclusion, which include words such as "therefore", "in conclusion", and "consequently"; indicators of contrast, which include words such as "but", "expect", and "not". The sentiment terminology is used to flag lexical semantic contrast. For example, "the flash worked poorly", as opposed to "the flash worked flawlessly" [176]. The user model is one tier of argumentative analysis. This model involves the properties of users that concern the quality of users reviews and users reactions responding in other reviews. The paper proposed four different subclasses of user properties consisting of: users parameters, which includes some aspects such as age, gender, and education; users context of use, which includes indoor, sport, and travel; users constraints that include cost, portability, and size; users quality expectations, which include color quality, reliability, and information density. The camera domain is concerned with terminology that relates to the camera. The main purpose of this tier is to identify the properties that are crucial to the users. The paper suggested three types of properties: Properties with ranges such as the number of megapixels; properties with binary values such as "has a flash"; multislotted properties such as the warranty. Overall, the focus in this paper is the identification of relevant textual units which can be parts of argumentation schemes and their attacks (counterargument) from digital camera reviews.

Arguments in other domains

Peldszus [118] proposed a study on argumentation mining that is concerned with a small corpus of German micro-texts contained authentic and explicit argumentations. The corpus

was created in controlled study using specific schemes. Then, the corpus was annotated for certain aspects of an argument in two different annotation studies using annotation guidelines. The first annotation study was carried out with 26 non-expert annotators, and the second study was carried out with three expert annotators. Both studies achieved scores of agreement reliability between 52% and 95%. The author then applied different machine learning classifiers (e.g., Naive Bayes and Support Vector Machines) to classify the argumentative components in this corpus. The author showed a comparison between classifiers results and demonstrated that the results were promising. However, since the corpus was small, these good results might not be the same in larger corpus.

Lawrence et al. [93] proposed an approach to analyze argumentation using text samples from 19th century philosophical book. The authors applied pre-processing steps (e.g., text segmentation and structure identification) to the texts to facilitate the automatic analysis of arguments. The segmentation process is employed to identify the texts into propositions using machine learning algorithm. The structure identification is used to determine the structure of an argument by calculating the score of a topical similarity between a proposition and its predecessor using the Euclidean metric. The proposition and its predecessor are connected if the score is below a set threshold and vice versa. A manual analysis for argumentative structures in the texts was carried out by an analyst to be used for training machine learning classifiers and evaluation purposes. The results showed that the automatic analysis of argumentation achieved an accuracy ranging from (11.6% to 20%) in case of identical match with the manual analysis.

2.3 Remarks

None of these previous approaches to automated argumentation analysis and mining provided a formal knowledge representation that could be used in detecting and recognizing argumentative elements. We believe that developing a formal representational framework based on verb semantics in procedural scientific discourse will enable a more in-depth analysis of argumentative elements in a computationally feasible manner.

In this chapter, we have examined the history of argumentation mining. We began by looking at historical aspects of argument theory, tracing back to Aristotle, Perelman and Olbrechts-Tyteca, Toulmin, and Tindale. We have also studied state-of-art works in argumentation research including argumentation schemes, and argumentation mining theories and techniques. We have found that argumentation mining gained traction in automatically analyzing legal documents. These documents usually have well-defined and clear argumentative structures. Therefore, analyzing these documents is the most obvious first

step in argumentation mining. As argumentation mining has gained more attention over the past few years, researchers have started looking into other text domains, such as scientific articles, which contain different form of argument structures than legal documents. To this end, researchers have proposed numerous argumentation schemes and machine learning algorithms to facilitate automatic mining in these articles. More recently, we have discovered a significant shift in interests. Many researchers have started looking into performing argumentation mining on data generated from social media such as online debates, online blogs, online news, and online forums. These data impose significant research efforts in analysis because they do not usually have well-defined argumentation structures or explicit argumentative components. Researchers have studied many aspects in advancing argumentation mining techniques in this domain. For example, some researchers attempted to crowd-source the task of argument annotations while others proposed to analyze the types of support statements for implicit claims, which are common in online textual data. Other researchers also proposed new argumentation systems or frameworks that facilitate automatic extraction of arguments from online resources. We find that recent publications mainly focused on online materials such as online debates, online user comments, and social media texts. We predict that the interests in analyzing argumentative structures of online texts will continue to expand.

Chapter 3

Rhetorical Moves for Biochemistry Articles

In this chapter, we present our proposed set of rhetorical moves to be detected in the analysis of biochemistry articles. Our discussion of motivation and related work in Chapters 1 and 2 reveals that: i) it is valuable to represent the argument structure of scientific articles ii) there is a proposed set of rhetorical moves for scientific articles that may be a useful starting point (the set proposed by Kanoksilapatham [80] which is a descriptive model, not explained in terms of computational processing) iii) insights have emerged on the benefits of producing an annotated corpus of biomedical articles, with work done by Green [62, 63] on how to guide this kind of annotation for the context of genetics articles; this enables manually tagging texts with their argument structure though Green's approach is not labelling rhetorical moves but instead identifying argumentation schemes

Our approach to argumentation analysis of biochemistry articles is to identify the set of rhetorical moves which could be considered in order to label the article with its overall structure. The method we used to derive this set was to study in detail a large set of articles (105), making an effort to tag each one with its underlying rhetorical moves. We began with an investigation of whether the series of moves proposed by Kanoksilapatham [80] served us well, when deriving the appropriate analysis of each article.

As will be explained below, we ended up deciding to make use of an abbreviated list of key possible rhetorical moves. This was done in order to simplify the processing, as we were aiming to support a computational analysis of the articles. Along the way, we learned about the most prevalent kinds of moves, and the importance of attaching somewhat different labels to some of the argument steps being identified than the ones proposed by Kanoksilapatham [80].

We also ended up observing the important role played by VERBS in the analysis. This led us to incorporate into our overall processing a step of identifying underlying verb frames and their semantic roles, which would deepen our final representation for the biochemistry article. The decisions we made for this component are described in Chapter 4. We also revisited the claim of Green that employing annotators would be helpful in identifying the argument structure of biomedical articles. In our case, this was to enable the annotators to tag rhetorical moves and semantic roles, yielding a rich tagged corpus of use in validating the algorithms we were proposing for automated analysis of the texts. We also go beyond Green's insights into how to perform argument annotation: delving into how best to combine the skills of domain experts with those of computational linguistics researchers. As will be explained in Chapter 5, we produce as contributions some guidelines for performing annotation, and some analysis of inter-annotator agreement.

3.1 Narrowing our Focus to the Methods Section of Biochemistry Texts

As discussed in Chapter 1, we decided to focus our attention on the Methods section of biochemistry articles, determining how to derive a representation of its argumentation structure. We present a sample biochemistry article in Appendix H and highlight its Methods section in Figure 3.1.

Our assumption[1] is that the Methods section of the article provides the central insights into the intended process of the experiment and as such its analysis would be especially useful to scientists concerned with finding out about the experiment (e.g. through a summary of content or question/answering – we revisit these possible end uses of the natural language analysis in Chapter 7). Also there is available a manual of standard experimental procedures [22, 133] that biochemists follow when they conduct their experiments. We note as well some research by Thompson [156] who analyzed and examined several articles published by a Nobel prize winning scientist, that also confirms that how inserting methodological justification or description of the method even in other sections such as "the result" section made the results more appealing and convincing. This claim leads to the conclusion that the most central claims of each article are in fact found within the Methods section.

[1]For future work, we could conduct a study to support the claim that this assumption is well-founded. See Section 8.3.9 for more details.

lowed by nucleophilic attack by the amino group of the condensing amino acids, with the formation of a peptide bond and the elimination of a phosphate group [7–12]. Inhibitors have been designed for MurD [9,13–15] and MurE [16,17]. Mechanistic and structural studies of the Mur enzymes and screening of these enzymes for inhibitors are severely hampered because of the lack of pathway intermediates [18]. The only commercially available substrate is UDP-*N*-acetylglucosamine, the substrate for MurA.

In this paper, we present the genomic analysis and organization of *murA* to *murF* genes in three distinct loci in the 6.3-Mb genome of *Pseudomonas aeruginosa*. We also present the purification of MurA, -B, -D, -E and -F in mg quantities at 99% homogeneity or more and the reconstruction of the enzymatic pathway for the biosynthesis of the pentapeptide peptidoglycan precursor. We reported the cloning and over-expression of *P. aeruginosa murC* elsewhere [6]. MurA to -F were combinatorially used to reconstruct the whole pathway in vitro and the final product was identified.

2. Materials and methods

2.1. General

All reagents were purchased from Sigma Aldrich (Oakville, ON, Canada) unless otherwise indicated. Buffer D was 20 mM potassium phosphate, 1 mM 2-mercaptoethanol, 0.1 mM $MgCl_2$, 15% (v/v) glycerol, pH 7.0 [19].

2.2. DNA manipulations, reagents and techniques

Restriction endonuclease and T4 ligase were obtained from New England Biolabs (Beverly, MA, USA). Agarose gel electrophoresis and plasmid DNA preparations were performed according to published procedures [20]. Recombinant plasmids containing *P. aeruginosa mur* genes were propagated in *Escherichia coli* NovaBlue (Novagen, Madison, WI, USA) prior to protein synthesis in *E. coli* BL21(λDE3) (Novagen).

2.3. Cloning of P. aeruginosa murA, –B, -D, -E and -F

Polymerase chain reaction (PCR) cloning was used to obtain MurA, -B, -D, -E and -F proteins with a His–Tag at their C-terminal. Upper and lower primers designed to contain appropriate restriction sites were designed as shown in Table 1. Five PCR reactions were performed with the upper and lower primers for each gene using genomic DNA of *P. aeruginosa* strain PAO1293 as the template. PCR conditions were optimized as follows: 30 cycles, denaturation at 95°C for 60 s, annealing at 55°C for 60 s, and extension at 72°C for 90 s, primers at 0.1 μM each, dNTPs (Amersham Pharmacia Biotech, Piscataway, NJ, USA) at 0.2 mM each, $MgCl_2$ at 2 mM, 5% dimethylsulfoxide (DMSO) in a final volume of 50 μl and adding 2.6 units of Expand high fidelity polymerase (Roche Diagnostics, Laval, QC, Canada) after Hot start of 7 min at 95°C. PCR products were purified using Qiaquick PCR purification kit (Qiagen, Chatsworth, CA, USA). Purified PCR products were digested with the restriction enzymes included in upper and lower primers and were cloned into the corresponding sites of the expression vectors pET30a and pET21 (Novagen) under the control of the bacteriophage T7 promoter.

2.4. DNA sequencing and computer analysis

Genomic analysis was done using data from the complete *P. aeruginosa* strain PAO1 sequence (www.pseudomonas.com) [21]. The sequences reported have the GenBank accession number AE004859 (*murA*), AE004723 (*murB*), AF110740 (*murC*), AY008276 (*murD*, -*E* and -*F*) and AE004091 (the complete genome). The DNA inserts in recombinant plasmids pMON3005, pMON3006, pMON3013, pMON3014 and pMON3009 (Table 1) were sequenced using T7 promoter primer and T7 terminator primer (Novagen). Sequence analyses were performed by the programs of Wisconsin Package Version 10.1, Genetics Computer Group (GCG), Madison, WI, USA.

2.5. Overproduction of P. aeruginosa MurA, -B, -D, -E and -F

The recombinant plasmids pMON3005, pMON3006, pMON3013, pMON3014 and pMON3009 (Table 1) were introduced into the *E. coli* host strain BL21(λDE3) (Novagen) by electroporation for expression of MurA, -B, -D, -E and -F respectively, with a His–Tag at their C-terminal. Overproduction was tested at two different incubation temperatures: 30 and 37°C, for three incubation periods: 3 h, 6 h and overnight (starting from the addition of isopropyl β-D-thiogalactopyranoside (IPTG)), using two different culture media: terrific broth and LB broth, and adding IPTG to two final concentrations: 0.5 mM and 1 mM (added after a cell density of $OD_{600\ nm} = 0.5$ was reached). Maximum protein yields in the soluble fractions were obtained after incubation for 6 h at 37°C using LB broth and adding IPTG to a final concentration of 1 mM. A small-scale overproduction pilot experiment using the optimized conditions showed that the soluble fractions of the proteins constitute 10%, 5%, 20%, 20% and 50% of their total protein fractions, respectively. Cultures were grown at 37°C in 1 l of LB broth containing 50 mg ml^{-1} kanamycin for MurA, -B and -F and 100 mg ml^{-1} ampicillin for MurD and -E, until a cell density of $OD_{600\ nm} = 0.5$ was reached. Cells were pelleted and resuspended in LB broth containing 1 mM IPTG. Cells were induced for 6 h, pelleted at 3000×*g* and frozen at −80°C [22].

Figure 3.1: The Methods section for the article in Appendix H

Move type	Definition
Description-of-method	Concerned with sentences that describe experimental events.
Appeal-to-authority	Concerned with sentences that discuss the use of well-established methods.
Background information	Concerned with all background information for the experimental events such as "method justification, comment, or observation, exclusion of data, approval of use of human tissue" as defined by Kanoksilapatham (2003).
Source-of-materials	Concerned with the use of certain biological materials in the experimental events.

Table 3.1: Rhetorical moves in the Methods sections of biochemistry articles

3.2 Gaining Insights into a Set of Rhetorical Moves to Model

The steps that we performed in order to derive the set of proposed rhetorical moves was actually quite laborious and time consuming. In the end, through personal inspection of a large number of sample texts and individual sentences, we converged on the set of rhetorical moves indicated in Table 3.1 One of our primary decisions was to combine some of the proposed moves of Kanoksilapatham [80], which is shown in Table 3.2, that is "Move4:Describing materials" with its steps "Step1:Listing materials", "Step2:Detailing the source of the materials" and "Step3:Providing the background of the materials" into the single category in our list namely, "Source-of-materials". This decision was mainly to come up with few rhetorical moves that are more comprehensive to facilitate the manual analysis of the biochemistry articles to lessen the burden for annotators with few choices. Another decision we made is to modify the definition of "Move5:Step1:Documenting established procedures" from simply referring to "an experimental process that is already established by previous researchers" to include any reference to an establish method, protocol or instrument. So, we name it "Appeal to Authority" to reflect that definition. We find that this move is the most important move because it showed how each decision that a scientist made in the lab is based on an established method that is widely accepted in the literature. We return to this as a possible label, in Section 3.2.1 below.

In Section 3.3 below we provide greater insights into the differences between the moves that we decided to model compared to those proposed by Kanoksilapatham [80]. We begin by listing the set of rhetorical moves proposed by Kanoksilapatham [80], for the Methods

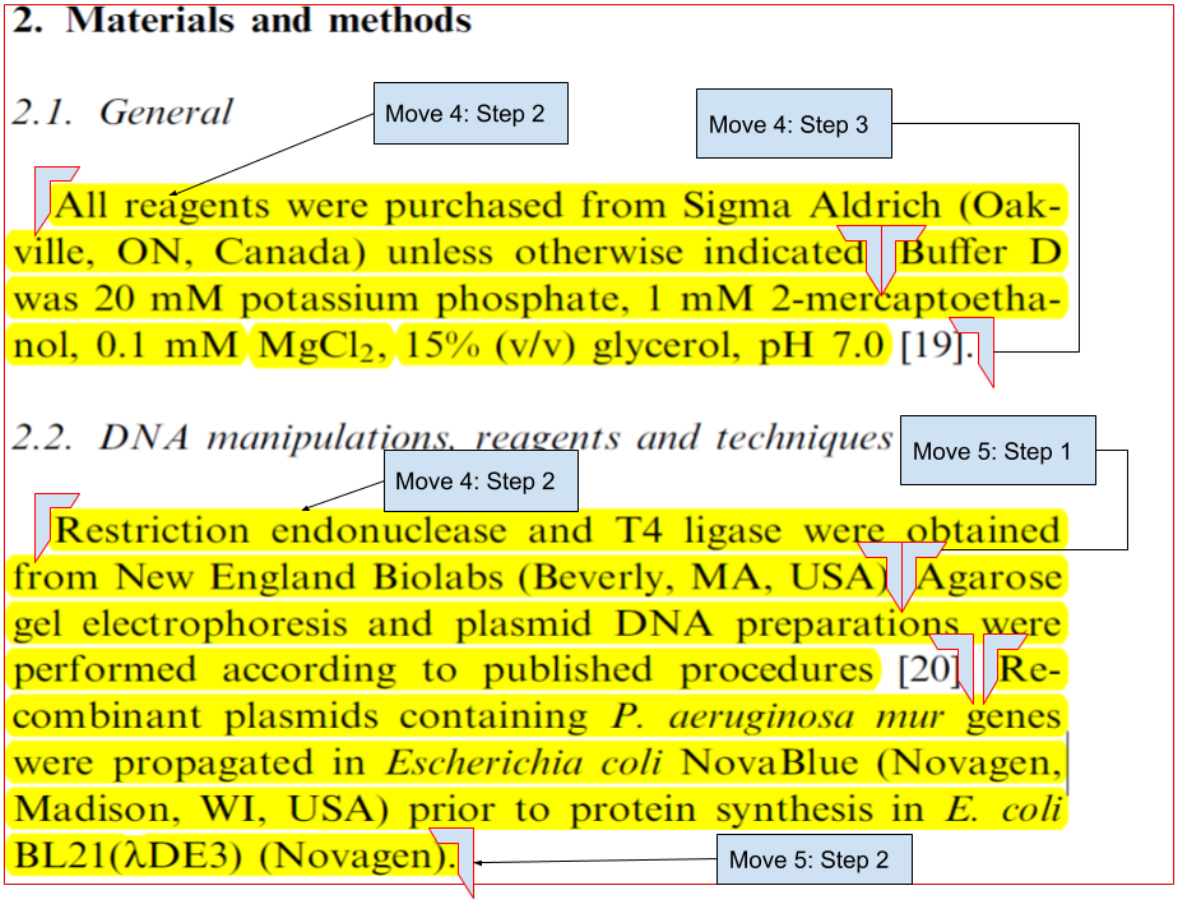

Figure 3.2: Annotation for the Methods section using Kanoksilapatham's [80] moves

section of biochemistry articles in Table 3.2.

3.2.1 Manual Tagging of Rhetorical Moves in a Corpus

The corpus used to explore the best choice for our set of rhetorical moves consisted of 105 articles. These articles were randomly selected from PubMed Central (PMC) biochemistry journal articles. Our first step was to examine each sentence, in turn, and to try to attach one the labels of Kanoksilapatham [80] to each one. Below we present a snippet for one example of our manual analysis that we examined and attempted to label, together with what we learned and what we ultimately decided about the best label to use. As can be seen in Figure 3.2, there are five sentences marked in the beginning and the end. The first

Move type	Definition
Move4:Describing materials	covers a wide variety of materials used in biochemistry ranging from natural substances, human/animal organs or tissues, to chemicals. Move 4 can be realized as one of the following steps:
Step1:Listing materials	explicitly itemizing materials or substances used in the study of the materials
Step2:Detailing the source of the materials	identifying how these items are obtained, such as, by purchase, as a gift, etc.
Step3:Providing the background of the materials	including the description, properties, or characteristics
Move5:Describing experimental procedures	indicates that biochemistry as a discipline is well established and its procedures, methods, and techniques are usually protocolized. This move has the following steps:
Step1:Documenting established procedures	recounts an experimental process that is already established by previous researchers. As a result of the standardization of experimental procedure, simple reference to the specific name of the method or procedure used to conduct research is adequate. Occasionally, certain procedures are unique or unorthodox for a particular study.
Step2:Detailing procedures	is used to provide detailed description of the procedures to enable future research replication.
Step3:Providing the background of the procedures	providing justification for the choice of technique or procedure, and comments or observations made during the experiment.
Move6:Detailing equipment (optional)	provides detailed information regarding the setting of the apparatus used for a particular task in an experiment
Move7:Describing statistical procedures (optional)	No definition was provided in [80]

Table 3.2: Kanoksilapatham's rhetorical moves in the Methods section of biochemistry articles [80]

sentence under section "2.1 General" is labelled as "Move4:Step2" according to Kanoksilapatham's moves [80] which basically listed the source of materials used. The second sentence is mainly to provide the background of the material used "Buffer D", so it should be labelled as "Move4:Step3". The remaining sentences under section "2.2 DNA manipulations, reagents, and techniques" in Figure 3.2 are labelled Move4:Step2, Move5:Step1, and Move5:Step2, respectively. We found that most of the moves that describe materials appeared at the beginning of the Methods section while moves that convey the description of experimental processes or justification of using certain biological items or instruments appeared later in the Methods section. Another important aspect from our analysis is that each move spans over a sentence which contains a main verb that characterizes that move. For example, the verb "purchased" in the first sentence (as seen in Figure 3.2) conveys the idea of obtaining something by means of paying. As another instance, the verb "performed" and the cue phrase such as "according to" in the fourth sentence in Figure 3.2 really communicate the idea of the use of an established experimental procedure. This type of observation led us to focus on analyzing verbs in biochemistry articles. We asked what are the most frequent verbs in the biochemistry articles? Are there domain specific verbs? Are these verbs associated more with certain moves than others? So we did an initial analysis by looking into how frequent these verbs were in the dataset we described earlier. We have used pdf and xml formats for the articles. The former format was mainly to analyze the article manually and the latter was to perform automatic analysis. First each xml file is modified to include only the Methods section. All figures are omitted. Then, each file is run through a series of pre-processing methods starting from splitting texts into sentences and sentences to tokens using NLTK (sent tokenzie, and tokenzie) [16] part-of-speech (POS) tagging GENIA Tagger [165]. We began by answering the first question and extracted the most frequent verbs in the aforementioned dataset. Table 3.3 shows the top 44 verbs that are frequent. The GENIA tagger mistakenly labelled the adverb "according" and the adjective "corresponding" as one of the verbs. We also omitted "Be" verbs such as "was" and "were". As can be seen in Table 3.3, we found that certain verbs are domain specific such as "Transfect", "Resuspend", "Equilibrate", and "Centrifuge". These verbs characterize the sentences. For example:

Example 1 *"Following linearization with Bgl II, the constructs were* ***transfected*** *into the P19 cell line using Lipofectamine (Gibco-BRL) at a concentration of 25 μl per 100 ml of serum free medium."[103]*

The above example shows how a domain specific "procedural[2]" verb is used to describe a step of an experimental procedure and we can see that this step described here is a part of

[2]We will use this term throughout the thesis which refers to domain specific verbs. We note that the term "procedure" has been used as well in the work of Bogost [19].

a series of steps. Also there is detailed information that is attached to the procedural verb "transfected" such as the use of a reagent "Lipofectamine", which is a common reagent for transfection. So, considering this point and other points earlier, we found that it is very significant to examine procedural verbs to acquire a better understanding of the sentences and their rhetorical moves. We decided to dig deeper into analyzing procedural verbs and attempted to find a lexical resource that describes these procedural verbs in terms of their usage in the sentence and to determine what are the associated arguments for each verb. However, we only found verbs described in lexical resources such as VerbNet [137] that are general and not domain specific. This led us to consult a domain expert who has the knowledge for the usage of these verbs and the definition of these verbs in contexts of biochemistry, that a working biochemist would use to describe experimental steps in a procedure. In Chapter 4, we describe our development of frame semantics based on the procedural verbs.

Verb	Frequency
Use	1830 times
Describe	609 times
Contain	569 times
Perform	432 times
Incubate	354 times
Purify	306 times
Determine	278 times
Add	271 times
Wash	261 times
Obtain	221 times
Prepare	201 times
Follow	261 times
Grow	150 times
Measure	149 times
Collect	148 times
Carry	143 times
Analyze	129 times
Supplement	119 times
Resuspend	118 times
Remove	115 times
Centrifuge	108 times
Dilute	102 times

Continued on next page

Continued from previous page

Verb	Frequency
Harvest	98 times
Clone	97 times
Express	95 times
Calculate	93 times
Store	93 times
Result	91 times
Purchase	90 times
Transfer	80 times
Transfect	78 times
Elute	77 times
Amplify	76 times
Apply	76 times
Extract	74 times
Indicate	71 times
Equilibrate	71 times
Assay	70 times
Separate	69 times
Produce	69 times
Load	69 times
Digest	68 times
Isolate	67 times
Make	66 times

Table 3.3: Top 45 verbs in the Methods sections from our 105 Biochemistry Articles dataset

Another important aspect about the aforementioned verbs in Table 3.3 is that most of the verbs with high frequency are in fact general ("non-technical") verbs which have been used in domain specific senses such as "use", "describe", "add" and "wash". So, we were attracted to discover how these non-technical verbs are used in this domain since these verbs are potential candidates to learn more about the constructions of rhetorical moves. Do they appear in sentences as main verbs or auxiliary ones? So, we analyzed most non-technical verbs in Table 3.3. We began by listing all sentences where a particular verb occurred and then examined each part of these sentences (e.g., words following the verbs). We found that there are common patterns in these words for every verb. For example, Table 3.4 and Table 3.5 show the frequency of several of these words for the verbs "used"

Words following the verb "used"	Frequency
The verb "used" followed by "for"	69 times
The verb "used" followed by "to"	65 times
The verb "used" followed by "in"	53 times
The verb "used" followed by "as"	32 times
The verb "used" followed by "throughout"	6 times
The verb "used" followed by "at"	6 times
The verb "used" followed by "with"	4 times
The verb "used" followed by "immediately"	2 times
The verb "used" followed by "and"	2 times
The verb "used" followed by miscellaneous items	39 times

Table 3.4: Patterns for words following the verb "used" from our 105 Biochemistry Articles dataset

and "washed". For example, various prepositions with high frequency such as "for" and "as" directly followed the verb "use". Let us examine the following sentences from our dataset:

- "The antiserum to Xenopus 20S proteasome and a monoclonal antibody to goldfish 20S proteasome $\alpha 2$ subunit were prepared and used as previously described [7,38]."
- "Mutational sense or antisense primers were used in parallel PCR reactions with the appropriate antisense or sense cloning primer, with HA-S1P4 plasmid DNA as template."
- "It should be noted that we initially expressed S1P4 in RH7777 cells, which are unresponsive to S1P and LPA and have been commonly used for studies of Edg family receptors [20]."

In the above examples, following our definitions of rhetorical moves, one could label the first sentence as the rhetorical move "Appeal to authority" since the sentence refers to the preparation and usage of two biological items namely "The antiserum" and "a monoclonal antibody" according to a method refereed in two references "[7, 38]". The second sentence should be labeled as "Description of the method" since it describes a step of experimental procedure of using certain "primers". The last sentence should be labeled as "Background information" because it explains the justification of why certain type of cells were selected instead of others.

Words following the verb "wash"	Frequency
The verb"wash" followed by number of times that a wash is done (e.g., once, twice, and three)	47 times
The verb "wash" followed by "with"	31 times
The verb "wash" followed by "in"	16 times
The verb "wash" followed by "and"	13 times
The verb "wash" followed by "exensively"	5 times
The verb "wash" followed by "as"	4 times
The verb "wash" followed by "at"	2 times
The verb "wash" followed by miscellaneous items	7 times

Table 3.5: Patterns for words following the verb"wash" from our 105 Biochemistry Articles dataset

This gave us an insight to look into more detail how the construction of attachments of non-technical verbs could contribute to the classification of rhetorical moves. Are verb-arguments for non-technical verbs the key to facilitate the classification of a sentence into the proper rhetorical move? Based on our analysis of different verb attachments and the association of certain verb-arguments earlier, we concluded that there is a need for a knowledge representation based on frame semantics to grasp and acquire more in depth understanding of sentences and their moves. We will shed some light on this aspect further in Chapter 4.

3.3 Reflection on Our Proposed Set of Rhetorical Moves

A couple of key distinctions are apparent between our set of rhetorical moves and those of Kanoksilapatham [80], as follows

- We consider the main frequent moves only (e.g., description of method and source of materials) while Kanoksilapatham's model includes moves that are less frequent in addition to the main ones.
- Our list of moves is designated for the Methods section only where Kanoksilapatham's moves cover all of the sections of biochemistry articles.

- Our list of moves is more comprehensive, meaning that one move category includes various steps where Kanoksilapatham's moves are fine grained, including steps for each move. Our decision was very critical and it has some advantages and disadvantages. One of the advantages we described earlier in this chapter is lessening the burden for annotators with fewer choices while some of the disadvantages will be revealed later in Chapter 5.

As will be explained in Chapter 5, when our expert annotators were asked to tag texts with their rhetorical moves, it was at times difficult to see complete agreement (e.g. whether Description of Method or Appeal to Authority was appropriate). We return to reflect on why this might be the case in Chapter 8, where we mention observations of Kanoksilapatham's model and possible future work to adjust our design decisions. At the end of the day, we developed a core set of moves that could be put in front of annotators with some ease and achieved successful annotation of a very large corpus of articles. As will be explained in Chapter 5, we moved on to examine almost 3500 different articles obtained from the top ten journals in biochemistry for a more in-depth study, during the annotation phase. A significant set of guidelines for our annotators emerged (see Appendix A).

Chapter 4

Semantic Roles

This chapter describes the development of our procedural rhetorical verb centric frame semantics as a knowledge representation for analyzing rhetorical moves in the biochemistry articles. This is one of the most crucial aspect of our work because it contributes to the analysis of experimental procedural verbs in the sentence(s) level which will in turn aid in the overall analysis of experimental procedures. This also can lead to better understanding of the characteristics of rhetorical moves in texts. We proposed that the identification of semantic roles of procedural verbs as a first step toward identifying *rhetorical moves*, text segments that are rhetorical and perform specific communicative goals, in the Methods section. Based on a descriptive taxonomy of rhetorical moves structured around an IMRaD (Introduction, Methods, Results, and Discussion) structure, the foundational linguistic knowledge needed for a computationally feasible model of the rhetorical moves is described: *semantic roles*. Using the observation that the structure of scholarly writing in the laboratory-based experimental sciences closely follows the laboratory procedures, we focus on the procedural verbs in the Methods section. Our goal is to provide FrameNet and VerbNet-like information for the specialized domain of biochemistry. This chapter presents the semantic roles required to achieve this goal. Section 4.1 describes the Experimental Procedure Writing, while Section 4.1.1 and Section 4.1.2 both discuss common procedural verbs and their associated semantic roles, respectively. Then, a list of frame semantics for common procedural verbs are provided in Section 4.2. We also describe the process of developing a semantic role labelling system and how we use the developed SRL system in Sections 4.3 and 4.4, respectively. Finally, we conclude the chapter with remarks in Section 4.5.

4.1 Experimental Procedure-oriented Writing

Our research goal is to provide a computational model for Kanoksilapatham's descriptive rhetorical move taxonomy which has been stated previously in Chapter 3 and later in Chapter 7. Initially, our focus is on the Methods section of the taxonomy since this provides a description of the procedures followed in the experiment, and the analysis of the results of the experiment thereby giving a framework for analyzing the moves in the remainder of the article. Because the experimental process is procedural, the moves tend to follow the verbs describing the steps in the experimental process. In other words, argumentation structure and scientific method both consist of rhetorical moves and experimental process, respectively. When a scientist describes her/his method in the writing, it contains a list of experimental steps which are described by verbs (actions). These verbs evoke (initiate) the rhetorical moves in the writing. To understand the moves, we need information about the semantic roles associated with these procedural verbs. Two well known databases containing semantic role information, Framenet [10] and Verbnet [137], do not provide the information appropriate for the verbs found in this scientific domain. So, our purpose in this chapter, in the spirit of these two databases, is to introduce the semantic roles that we are proposing for this domain, some of which are the same as those normally found and some which are new and we suggest are required for this domain.

Scientific writing in the biochemistry domain has certain characteristics that make it ideal for our purposes. In this domain, experimental procedures describe the sequence of actions the biochemist performs to carry out an experiment to derive scientific conclusions, to demonstrate science experiments as can be seen in the experimental manuals (e.g., [22, 133]). Verbs play an essential role as indicators of these experimental procedures. These procedures can be viewed as corresponding to the elements of the scientific argumentation structure. For example, when examining a biological substance (e.g., a certain type of bacteria) in order to prove a hypothesis (e.g., this bacteria is correlated with a certain disease) the biochemist would perform a sequence of certain procedures to arrive at a conclusion. Essentially, biochemists create an argumentation framework through the scientific methodology they follow—how they perform their experiments is how they argue. We can observe that this genre—biochemistry articles—is procedure-oriented since the scientific procedures that are described are parallel to the scientific argumentation in the text.

For example:

Example 2 *"Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." [47].*

In this example, the verbs "washed", "harvested", "separated", and "analyzed" are used to illustrate the procedure steps in sequential order. Such an experiment can be reproduced if one follows these steps.

Minsky defined *frame* as "a data-structure representing a stereotyped situation" [107], with frames having a header, slots and slot fillers. Fillmore [49] introduced the notion of frame semantics as a theory of meaning. A *semantic frame* is defined by Fillmore as "any coherent individuatable perception, memory, experience, action or object"[1] [50]. In other words, these are coherently world events or experiences. In our case, we develop frame semantics at the verb level so that our headers are verbs and our slots are semantic roles, filled by the words which represent these roles. For example, to understand the word "buy", one would access the knowledge contained in the commercial transaction frame which includes words such as the person who buys the goods (buyer), the goods that are being sold (goods), the person who sells the goods (seller), and the currency that the buyer and seller agree on (money).

Motivated by Fillmore's theory of frame semantics, FrameNet [10] was developed to create an online lexical resource for English. This framework includes more than 170,000 manually annotated sentences and 10,000 words. The computational linguistic community has been attracted to the concept of frame semantics and developed computational resources using this concept, such as VerbNet [137], an on-line verb lexicon for English and PropBank [115], an annotated corpus with basic semantic propositions. Following the notion of VerbNet, we propose to build a knowledge representatnvloon framework to analyze verbs in a procedurally-oriented genre. Our concept of verb-centric frame semantics is intended to address this gap by developing a computationally feasible knowledge representation, based on semantic roles, that will enable the analysis of rhetorical moves. Our hypothesis is that development of a frame-based knowledge representation can be based on the semantics of the verbs associated with these procedures. The knowledge representation used to represent the verbs and their semantic roles is frame-based.

The concept of frames originates with the work of Minksy [107] and refers to a representation with a header (in our case, the central verb) and slots (for us, the possible semantic roles) and their fillers. This representation can provide detailed knowledge for understanding these rhetorical moves. In other words, we propose that a *procedurally rhetorical verb-centric frame semantics* can be used to obtain a deeper analysis of sentence meaning in a computationally feasible manner.

Before moving to our description of the procedurally rhetorical verb-centric frame se-

[1]Fillmore developed this concept further in his later work, but this definition is sufficient for our purposes here. Our frames perhaps best resemble what he refers to as event schemata. We used the phrase experimental event scheme at times when discussing these frames.

mantics, we wish to inform the reader that Section 4.6 provides a more detailed discussion of the process that led to this knowledge representation.

4.1.1 Procedural Verbs

We have studied closely 39 procedural verbs in biochemistry articles. Some of these verbs were selected based on their frequency [2], and their importance in major experimental events in biochemistry procedures such as DNA purification, ligation and digestion while other verbs were suggested by the domain expert. These verbs include, but are not limited to, "Annelae", "Bind" , and "Biotinylated" etc. We have described their definitions and their usages in Appendix E. Table 4.1 and Table 4.2 shows two examples of the verbs. One of the important observation for some verbs (e.g., "bind") that we noted when working with domain experts is that many of the sentences in the biochemistry articles do not even describe what is happening in the binding; they are just describing how they measured binding. The sentences refer to binding as itself because they understand that that part is just what happens. If there were verbs describing specific protocols that emphasize binding as being important that would be more efficient than trying to understand all of the ways that binding is used at once. Examples of protocols where binding is important are: Affinity chromatography, screens for protein-ligand binding efficiency, and molecular markers. The specific ways in which binding occurs are literally infinite. The central dogma of biochemistry is that proteins bind only one (or a few) thing(s) specifically and facilitate a particular interaction (in the general case). Generally, binding can occur between each of the following as well as themselves: Atoms Molecules, Cofactors, Proteins, RNA, and DNA. The specific nature of the binding can differ vastly however in all these interactions.

4.1.2 Semantic Roles

We are focusing on procedural verbs with the associated semantic roles in this genre. Verbs evoke semantic roles in writing. Semantic roles provide salient pieces of information about experimental steps. Although lexical resources such as FrameNet and VerbNet provide syntactic and semantic frames for most common verbs in English, these resources do not provide such a knowledge for the aforementioned procedural verbs in Table E.1. We intended to create these frames for our list of verbs which require similar semantic roles that are normally found in general English use, and others which are required by this domain. We have developed our experimental event scheme for verb arguments which is

[2] our early analysis for the data set of 105 articles which was described in Chapter 3

Verb	Definition and usage
Bind	**Definition:** this is more of a chemistry term that means any sort of interaction between atoms or molecules that involves them coming together and sticking to one another. There are many different kinds of bonds/binding and some are stronger and some are weaker. The strongest bonds are ionic bonds, but those are almost never of interest in biochemistry or molecular biology. Covalent bonds are slightly weaker than ionic bonds and are what hold most organic molecules together. For example, in a glucose molecule there are a number of carbon, oxygen and hydrogen atoms. Those are all non-metals meaning they covalent bond to one another. There are weaker electrostatic interactions than that such as hydrogen bonds which are very important for proteinligand binding as often it is hydrogen bonds that form the active or binding site. Binding can be done actively by researchers for different reasons but it also occurs non-stop in nature, so at times, binding is the subject of study rather than the protocol (like in signaling pathways). Synonyms: bond, bonded. **Why is it done:** There are many reasons one would want to molecules to bind. You could put a molecular marker on a protein to track it, you might want to see a potential substrate drug bind to the protein of interest to inhibit it or adjust it slightly to try to optimize the binding, or you might be using it to purify a protein in a column. Binding is literally everywhere so it is difficult to identify all the protocols it might be a part of. You might do it to: proteins, ligands, enzymes, substrates, DNA, RNA, molecules, cofactors, beads, membranes. **How it is done:** By forming immunocomplexes, electrostatic interactions.

Table 4.1: Example 1 of common procedural verbs of biochemistry articles

Verb	Definition and usage
Digested	**Definition:** In molecular biology this means cut or cleaved (usually DNA). But it can also mean in certain contexts that a large end of DNA was destroyed or removed or that something was destroyed entirely. **Why is it done:** To ligate something into a plasmid, subcloning, molecular cloning, preparing a sample for something. You might do it to: DNA, RNA, Protein. (In order of what you will see most frequently). **How is it done:** with a specific restriction enzyme, at a particular cut site, from a certain 'end', at a particular temperature, for a certain amount of time, in a particular buffer, at a particular sequence/coding or non-coding region. Often digested will be followed by "with" which describes the restriction enzyme or enzymes responsible for performing the digestion.

Table 4.2: Example 2 of common procedural verbs of biochemistry articles

based on the inventory of semantic roles in VerbNet [137] and modified and added new semantic roles to define our scheme. Our experimental event scheme includes: *Theme*, *Patient*, *Agent*, *Location*,*Goal*, etc. The complete set of semantic roles and their definitions in our experimental event scheme is presented in Table 4.3. These semantic roles have been defined and examined carefully based on our analysis of a large number of experimental procedures with domain experts using a data set of 105 articles from PubMed biochemistry. As can be seen in Table 4.3, we have developed a sufficiently large set of roles that identifies the arguments of both verbs and nominalised verbs. This leads to increase on time spent and greater burden on labelling semantic roles by human annotators.

We have extended the VerbNet definition of the semantic role *Instrument* from simply describing "an object or force that comes in contact with an object and causes some change in them" [137] to include a variety of subcategories that correspond to various types of biological and man-made instruments that are used in a biochemistry laboratory (see Table 4.3). We have also proposed a new semantic role *protocol detail* that identifies certain types of information about experimental processes. *Time* and *Temperature* were proposed in [157]; however, *condition, buffer* and *cofactor* are new additions. The semantic role *Factitive* was also developed by [100]. *Goal* was proposed in the literature [51, 90, 98]. We have extended the definition of the semantic role *Goal* to include two categories: *Physical* and *Purpose*. The former is similar to the one proposed in the literature and the latter is newly proposed and required in this domain based on the annotators' feedback.

Semantic role	Definition
Agent	"Generally a human or an animate subject. Used mostly as a volitional agent, but also used in VerbNet for internally controlled subjects such as forces and machines" [3].
Patient	"used for participants that are undergoing a process or that have been affected in some way" [4].
Theme	"used for participants in a location or undergoing a change of location" [5].
Goal:Physical	Identifies a thing toward which an action is directed or place to which something moves [6].
Goal:Purpose	Identifies the stated purpose in a sentence for doing certain actions.
Factitive	"A referent that results from the action or state identified by a verb" [7].
Location	The physical place where the experiments took place.
Protocol-Detail:Time	Identifies the time or a duration of an experimental process.
Protocol-Detail:Temperature	Identifies the temperature of an experimental process.
Protocol-Detail:Condition	Identifies the condition of how an experimental process being carried out (e.g., under rotation).
Protocol-Detail:Repetition	Identifies the number of times that an experimental process being repeated.
Protocol-Detail:Buffer	Identifies the buffer that was used in an experimental process.
Protocol-Detail:Cofactor	Identifies the cofactor that was used in an experimental process.

Continued on next page

[3]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[4]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[5]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[6]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[7]http://www.glossary.sil.org/term/factitive-semantic-role

Continued from previous page

Semantic role	Definition
Instrument:Change	Describes an object or protocol that can change another object(s). This role corresponds closely with the VerbNet project instrument semantic role which describes something "used to describe objects (or forces) that come in contact with an object and cause some change in them".
Instrument:Measure	Describes an object or protocol that can measure another object(s).
Instrument:Observe	Describes an object which can be used to observe another object(s).
Instrument:Maintain	Describes an object or protocol which can be used to maintain the state of object(s).
Instrument:Catalyst	Describes an object that can be used as a catalytic "facilitator" for an experimental event to occur.
Instrument:Reference	Refers to a method or protocol being used.
Instrument:Mathematical	Describes a mathematical or computational instrument (e.g., simulation, algorithm, equation, and the use of software).

Table 4.3: Semantic roles in the annotation scheme of our experimental event

4.2 Frame Semantics

A key aspect of our hypothesis is that development of a frame-based knowledge representation can be based on the semantics of the verbs associated with these procedures. This representation can provide detailed knowledge for understanding these rhetorical moves. We have created frames for these aforementioned procedural verbs in Table E.1. These frames were created carefully following the guidelines provided by the VerbNet [137] and Propbank [9]. Domain experts, fluent and native speakers, in biochemistry were involved in the process of creating these frames. We have used the annotated data set (Gold standard) by human annotators to assess our decision in creating these frames. We tried to be comprehensive in our selection of the examples for each frame from the data set and select roles which are semantically required and appeared to occur frequently. On the one hand,

FRAMES for digest-121
NP V PP
[NP Array-generated oligos] [VP were digested] [PP with] [NP restriction enzymes] ([NP Not1] and [NP EcoR1]
Example: “Array-generated oligos were digested with restriction enzymes (Not1 and EcoR1)”
Syntax: PATIENT V INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V ADVP PP VP
[NP fractions] [PP of] [NP interest] [VP containing] [NP NS DNA] [VP were digested] [ADVP twice] [PP with] [NP λ-Exo] [VP to eliminate] [NP contaminating DNA]
Example: “fractions of interest containing NS DNA were digested twice with λ-Exo to eliminate contaminating DNA.”
Syntax: PATIENT V REPETITION INSTRUMENT GOAL
Semantics: MANNER (DURING(E), PATIENT)
PP NP V PP PP
[ADVP Moreover] ,[PP as] [NP a negative control] [PP in] [NP the above study] , [NP a large amount] [PP of] [NP total fragmented DNA] ([NP 150 µg]) [VP was digested] [PP with] [NP λ-Exo] [PP in] [NP strong limiting conditions] ([NP 0.7 units] [PP of] [NP λ-exo /µg] [PP of] [NP DNA])
Example: “Moreover, as a negative control in the above study, a large amount of total fragmented DNA (150 µg) was digested with λ-Exo in strong limiting conditions (0.7 units of λ-exo /µg of DNA).”
Syntax: GOAL PATIENT V INSTRUMENT CONDITION
Semantics: MANNER (DURING(E), PATIENT)
NP V VP PP PP PP
[NP Forty micrograms] [NP total RNA] [VP was digested] [VP using] [NP 0.1 U nuclease P1] ([NP Yamasa Corporation]) [PP in] [NP 25 mM NH4OAc] ([NP pH 5.3]) [PP at] [NP 37 °C] [PP for] [NP 1 h]
Example: “Forty micrograms total RNA was digested using 0.1 U nuclease P1 (Yamasa Corporation) in 25 mM NH4OAc (pH 5.3) at 37 °C for 1 h.”
Syntax: PATIENT V INSTRUMENT BUFFER TEMP TIME
Semantics: MANNER (DURING(E), PATIENT)
NP V PP VP PP
[NP DNA] [VP was digested] [PP with] [NP HindIII restriction enzyme] [VP leaving] [NP an overhang] [NP that] [VP is filled] [PRT in] [PP by] [NP biotinylated dCTP]
Example: “DNA was digested with HindIII restriction enzyme leaving an overhang that is filled in by biotinylated dCTP.”
Syntax: PATIENT V INSTRUMENT FACTITIVE CONDITION
Semantics: MANNER (DURING(E), PATIENT)

Figure 4.1: The frame for the verb *digest*

generally, some verbs have more than one framesets that correspond to various senses such the case in the propbank project [9]. On the other hand, since we are interested in verbs that are tailored to this specific domain, experimental procedures, we only consider verb sense, definition, described in Table E.1. We only created VerbNet-like frame semantics for the verbs which have instances in our data set because not all verbs in our list, which are shown in Table E.1, appeared in our data set. Figure 4.1 shows an example from our list of syntactic frames for the verb *digest*; the complete list of verb frames is added in Appendix F. These syntactic frames have different argument structure alterations, such as unspecified argument (e.g., time, and temperature). In addition, verb frames contain semantic predicates, as can be seen in Figure 4.1. "These predicates are expressed as a conjunction of boolean semantic predicates such as 'motion,' 'contact,' or 'cause.' Each predicate is associated with an event variable E that allows predicates to specify when in the event the predicate is true (start(E) for the preparatory stage, during(E) for the culmination stage, and end(E) for the consequent stage)" [84]. In our case, our list of verbs denotes activities or processes. So, following VerbNet guidelines [84, 137], our verbs have semantic predicates that only refer to 'motion' and 'manner', which depends on the activity of a particular verb, associated with the during(E) stage of the event.

As we discussed earlier, some information could be implicit such as time and temperature. This information could be obtained from the ontology of experimental procedures. We described in Chapter 6 an ontology for one experimental procedure and it demonstrated how powerful its use is in extracting relevant missing (aka, implicit) information. These components, the use of frame semantics and ontologies, are very important steps in our proposed framework. Another important aspect is that some pieces of implicit information can be inferred from the standard laboratory conditions meaning that there are standard default values that a biochemist would infer for that missing piece of information (e.g., if temperature is not mentioned in the text, the default temperature is 23°C)[8]. As a side note, we found that some verbs (e.g., equilibrate) when appearing in our dataset sentences, usually come after the noun phrase (e.g., Theme or Patient) without the use of a "Be" verb (e.g., was and were). We assumed that this could be a writing style.

4.3 Semantic Role Labelling

The main goal of semantic role labelling (SRL) system is to recognize the predicate-argument structure of a sentence by finding relevant information in the sentence such

[8]Some labs have restrictions for specific temperature (e.g., 23-24°C) and humidity (e.g., at 50%) if these conditions drop below the recommended ones, the chances of having contaminates will be higher.

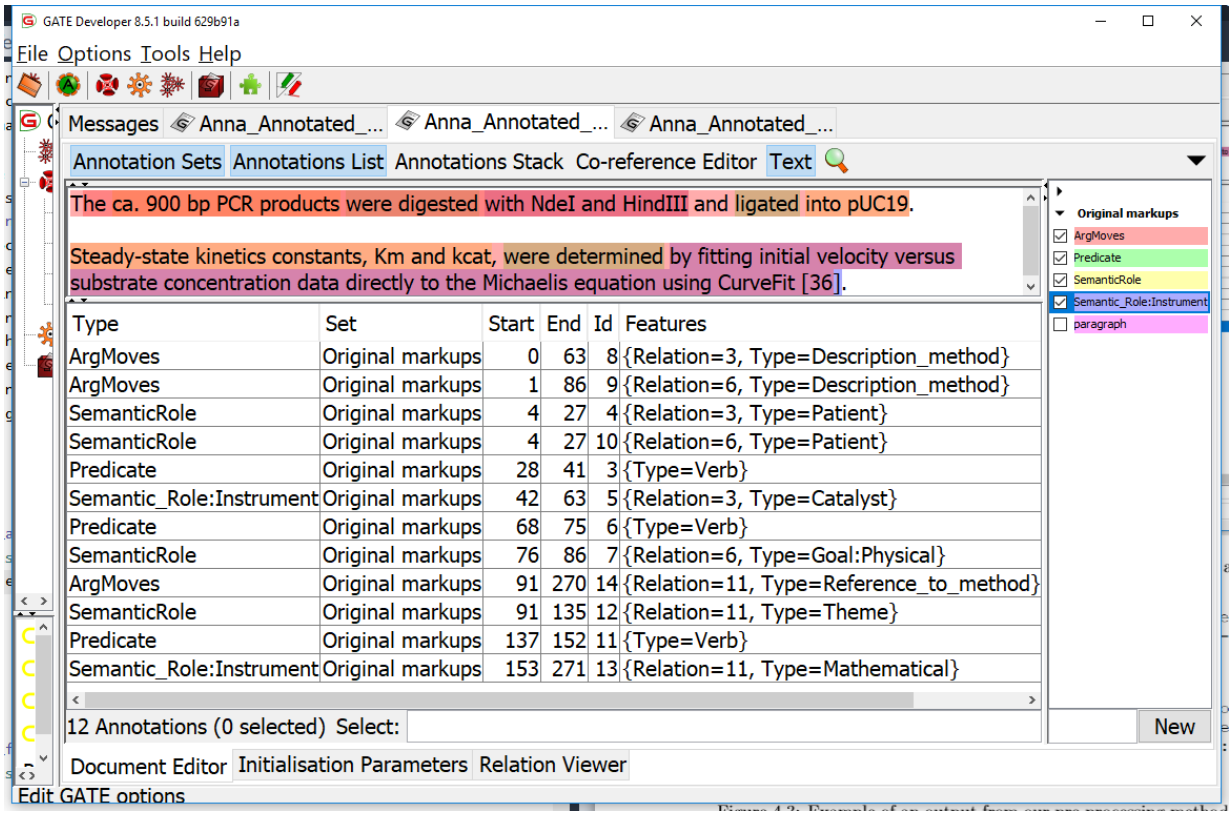

Figure 4.2: Example of our dataset annotation done by annotators

```
6 The ca. 900 bp PCR products were digested with NdeI and HindIII |||
B-Patient I-Patient I-Patient I-Patient I-Patient I-Patient B-V I-V B-
Instrument:Catalyst I-Instrument:Catalyst I-Instrument:Catalyst I-
Instrument:Catalyst O

6 The ca. 900 bp PCR products ligated into pUC19. ||| B-Patient I-
Patient I-Patient I-Patient I-Patient I-Patient B-V B-Goal:Physical I-
Goal:Physical O

6 Steady-state kinetics constants, Km and kcat, were determined by
fitting initial velocity versus substrate concentration data directly
to the Michaelis equation using CurveFit [36]. |||  B-Patient I-
Patient I-Patient I-Patient I-Patient I-Patient B-V I-V B-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical I-
Instrument:Mathematical I-Instrument:Mathematical O
```

Figure 4.3: Example of an output from our pre-processing method to prepare the training data in BIO2 tagging. Each tag indicates the label for each token (word) in the sentence.

as 'What happened", "Where it happened" and "When it happened". Several SRL systems were proposed in the last decade ranging from syntax-based approaches such as [122] to end-to-end deep learning approaches [67] which does not require syntactic information.

4.3.1 Pre-processing Step for Our Model Learning

In this work, as Anthony J. D'Angelo said "don't reinvent the wheel just realign it". We have used the SRL system developed by [67], a deep neural network model, for few reasons. First, it outperformed all state-of-art SRL systems such as [52, 150, 180] on datasets of CONLL2005, which consists of the Wall Street Journal sections 02-21 and 24 of the Penn Treebank II collection annotated with syntactic and semantic information [28], and CONLL2012, a large corpus contained various types of texts ranging from news

articles to to talk shows in three languages Arabic, Chinese and English annotated with syntactic and semantic information [121, 174][9]. Second, it is publicly available which is a huge advantage. Essentially, we have pre-processed our dataset which contained the human-based annotation that we described in detail in Chapter 5. Since our annotated corpus is in xml format (Figure 4.2 shows an example of our annotated dataset performed using GATE), we need to pre-process it to be suitable for training the neural network model which helps to validate our solution. We have developed an algorithm that takes an annotation xml file from our dataset, as shown in Appendix G, and transforms it to the required format. This process is done to all files in our dataset, for the model. So, each sentence should be in one line with the positional index of a predicate in that sentence followed by a separator and the tags of semantic role labelling, which is derived from our annotated dataset, for each argument in these sentences in BIO2 encoding[10], which is simply tagging tokens in a chunk as shown in Figure 4.3. Figure 4.3 showed three lines and each line starts with a number (i.e., the verb index in a sentence) followed by a sentence followed by a '|||' mark and ends with BIO2 tagging for the sentence before the '|||' mark. Basically, Appendix G and Figure 4.3 show an example of the input and the output respectively from our pre-processing stage.

4.3.2 Model Learning

We developed a deep Recurrent Neural Network (RNN) model that uses a highway BiLSTM (Bidirectional Long Short-Term Memory) architecture with constrained decoding as proposed in [67]. RNN has great potential in modeling sequence problems due to the fact that RNN maintains a form of memory of previous inputs which enables the use of past information in the network. However, the range of the past information is limited because gradient parameters can handle processing long sequences and they may vanish or explode. This phenomenon is called in practice "the problem of vanishing gradient". So, LSTM, which is a type of RNN, was proposed to overcome this shortcoming by using memory blocks that contain cells with three gates: namely the input gate, forget gate and output gate (see Figure 4.4). These gates enable the cells to maintain information for a longer time compared to regular RNN [60]. This model applied several key components: Highway connections which introduce gated directed connections between memory cells and other neighbour layers to allow free flow of information across different layers to diminish the

[9]Note that we don't use HMM or CRF for machine learning because of the good performance of this model, explained above.

[10]The **B**eginning **I**nside **O**utside (BIO) format is a tagging format that is used to tag tokens in a chunking task [123].

gradient vanishing problem [178], and a recurrent dropout which temporarily eliminates units in a neural network to reduce the over-fitting problem [147].

We have trained our model on the developed annotated dataset. We have divided our annotated dataset into training (50%), development (25%), and testing set (25%). Our dataset contains 16847 sentences. However, we only were able to use 8778 sentences for (training/testing/development, namely (4389, 2194, 2194)) since the maximum length for each instance should be less or equal to 100 tokens as suggested by [67]. We will talk about this issue in Chapter 8. Following the configurations by He et al. [67], we have trained our model which consists of eight BiLSTM layers, including four forward LSTMs and four reversed LSTMs. The dimensional hidden units are 300, and a softmax layer is used for predicting the output distribution. The training time for the model with 8 layers was one week. Once trained we have an SRL model that is used to give semantic role labels for a verb in a sentence. We have also trained a model for predicates. This model is trained on the same training data with BIO2 tagging except that only the beginning of a predicate is tagged by 'V' and everything else is tagged with 'O'. This leads to an important issue to deal with when training the model. Almost 90% of the sentences in our dataset are in passive voice meaning that almost always "Be" verbs such as "was and were" precedes the main verb. Annotators were trained to label the "Be" verbs with the main verbs. For example:
"The cells **were washed** twice ..."

So, "were washed" is labeled as a "predicate" for that sentence. However, "were" is tagged in BIO2 as "B-V" and "washed" as "I-V". Therefore, we decided to train a model with a training set where only the main verb is tagged by 'V' which in the above example is "washed".

4.3.3 Results and Discussion

We used the data set that has been annotated by domain experts, which described in Chapter 5 (see section 5.2). The model was trained on a training set consisting of 4389 sentences and a development set of 2194 sentences. After the model was trained, we have tested our model using the test set which consists of 2194 sentences. We have used K-fold cross-validation to evaluate the performance of our model. Essentially, k-fold cross validation is used to evaluate machine learning models by re-sampling the dataset into k-folds. In our case, we select k to be 10. So the data set is shuffled randomly and using a seed of the pseudo random number generator to ensure we have a unique shuffled dataset among the 10-folds. Also since we have used a deep neural network model, this extracts features from the hidden layers automatically. This is one of the advantages of using a

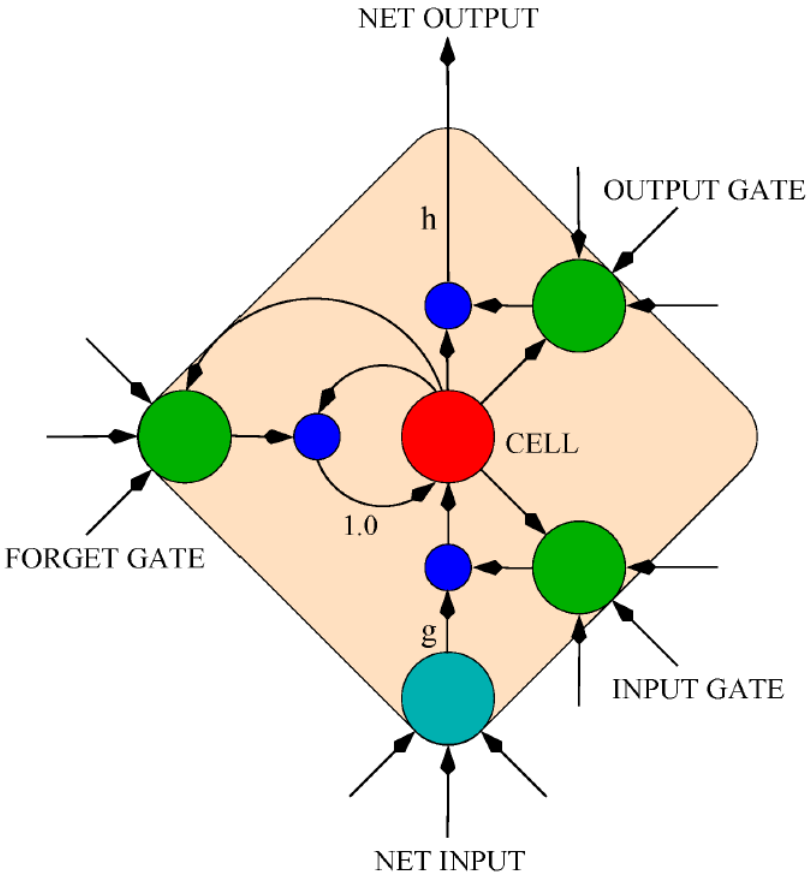

Figure 4.4: LSTM memory block with one cell (from Graves et al. [60])

deep neural network because it does not require any feature engineering manually. The results in Table 4.4 show the performance of our model. We have achieved an overall 48.54% F1 score for over all semantic roles of 10-fold cross validation. Table 4.4 shows promising results given that our dataset is small. "Correct" means the model was able to correctly label verb arguments into the proper semantic role category. "Missed" refers to the model missed labelling the arguments. "Excess" indicates that the model labels an argument incorrectly. As can be seen in Table 4.4, "Agent", on the one hand, achieved the highest F1 score since this role is one of easiest role to predict; however, we only have few instances of this role since sentences in the method section tend to be in passive voice. On the other hand, the model performed poorly in predicating the role "Cofactor". This is due to several aspects: First, we noticed that there are disagreements among annotators when we performed the inter-annotator agreement scores earlier in Chapter 5. We also noticed that when we merged all protocol detail sub-categories into one main category the kappa score improved.

4.4 Our Developed SRL System

We have described how we trained our model in the previous section. Now, we describe how to use our trained model in our overall framework. We have developed an algorithm that takes a set of sentences as an input and produces a file that contains the predicate(s) and semantic roles for each input sentence. Our SRL system is part of our computational framework which is described in Chapter 7. So, we are going to describe our computational model that starts from the input and ends with producing an output. Our SRL system includes several pre-processing steps before the predication step.

4.4.1 Pre-processing Step for Our SRL Prediction

First we assume that the input is in text format which contains only the Method section. Then, each file is run through a series of pre-processing methods: Sentence tokenization, and then the text is split so that each sentence is on one line. Then each sentence is tokenized using a word tokenizer [16]. The tokenization is mainly used to obtain the part-of-speech (POS) tagging using the GENIA Tagger[165], which is a well-known tagger that has been trained on biomedical data. The POS tagging is an important step to identify the verb(s) in a given sentence. For example, let us use the first sentence in Figure 4.4:

"the ca. 900 bp PCR products were digested with NdeI and HindIII"

and use it as an input for the GENIA tagger. It will produce the following:

	Correct	Excess	Missed	Precision	Recall	F1
Overall	2589	2719	2770	48.78%	48.31%	48.54%
Agent	104	10	16	90.72%	86.28%	88.42%
Buffer	25	48	48	33.94%	34.28%	33.92%
Theme	960	860	731	52.77%	56.76%	54.67%
Patient	294	299	355	49.65%	45.25%	47.23%
Factitive	79	87	129	47.87%	38.03%	42.32%
Goal:Physical	33	72	83	31.53%	28.44%	29.80%
Goal:Purpose	103	94	92	52.46%	52.82%	52.57%
Location	5	13	18	31.46%	23.61%	25.74%
Instrument:Catalyst	2	10	15	20.43%	15.83%	17.58%
Instrument:Change	100	171	166	37.43%	37.76%	37.32%
Instrument:Maintain	12	43	52	23.23%	19.84%	21.21%
Instrument:Mathematical	82	122	114	40.49%	42.03%	41.08%
Instrument:Measure	28	61	66	31.60%	29.72%	30.33%
Instrument:Observe	12	22	28	37.27%	31.27%	33.69%
Instrument:Reference	211	122	112	63.35%	65.32%	64.26%
Repetition	28	20	32	59.60%	46.95%	51.35%
Temp	93	32	40	74.8%	69.89%	72.09%
Cofactor	8	40	62	16.22%	11.21%	12.88%
Condition	286	528	532	35.28%	35.01%	35.11%
Time	116	58	72	66.97%	61.74%	64.17%
Verb	2170	22	23	98.97%	98.92%	98.94%

Table 4.4: Precision, recall, and F1 scores of semantic role labeling for the average of 10-fold cross validation

[the, 'DT'] [ca, 'NN'] . [900, 'CD'] [bp, 'NN'] [PCR, 'NN'] [products, 'NNS'] [were, 'VBD'] [digested, 'VBN'] [with, 'IN'] [NdeI,'NN'] [and, 'CC'] [HindIII, 'NN']

Each sentence is also chunk parsed by the GENIA parser [165] to find the chunk parse of a sentence. Then, our sentence is fed to the pre-trained SRL model. The verbs (predicates) in the sentence have their semantic roles predicted by the SRL. These labels are attached to the noun phrases and the prepositional phrases in that sentence. For example,

Predicate: digested
Theme: the ca . 900 bp PCR products were
V: digested
Instrument:Catalyst: with

Instrument:Catalyst: NdeI and

This is one stage from our overall framework, which will be described in Chapter 7.

The predicate has been identified in this example: "digested". The second stage is to identify the frames of this verb from our pre-defined set of syntactic frames. Since we have the chunk parse tree for each sentence from the previous stage (i.e., the output from the Genia parser), the parse of a sentence is checked using the pre-defined regular expression rules from our frames for a given predicate (e.g., "NN V PP"). In this case, the verb is "digest", so a set of rules is generated. So, we have 5 rules for the verb "digest". The parsed sentence is checked against each rule to find a match. Once a match is found the syntax frame is compared with the labeled sentence from the shallow parsing step (the output of the SRL). If they match, then the next step is the rhetorical move labeling stage. For example, let's assume the following sentence:

"The resulting ca. 900 bp piece was gel purified"

The frame for the verb "purify" requires "NP V PP.instrument". However, in the above example only "NP V" is given, this is a partial match. The verb phrase "gel purified" will trigger an ontology query using SPARQL pre-defined commands that are associated with this verb to find the missing "instrument" which is in this case "electrophoresis".

4.5 Remarks

In this chapter, we have presented the semantic roles that we have suggested to be necessary for this scientific domain and which will be used in our annotation scheme. This Experimental Event Scheme, which is based on the proposed semantic roles, is the first step towards developing an automated rhetorical moves analysis (discussed in more detail in Chapter 7). We have developed a small set of frames for some verbs (e.g., "wash") based on the manually analyzed data. We also aim to extend the VerbNet project [137] by providing syntactic and semantic frames for these aforementioned procedural verbs. In future work, we aim to develop a larger set of frames for the most frequent procedural verbs in biochemistry. Our experience with annotating the biochemistry articles with our experts, we recognized that not all of the information needed to interpret the move structure is available in the text. What is needed is an ontology that captures the knowledge that a working biochemist would have regarding biochemistry experimental procedures, especially the sequence of events that are normally undertaken in these laboratory procedures. We describe building such an ontology in Chapter 6.

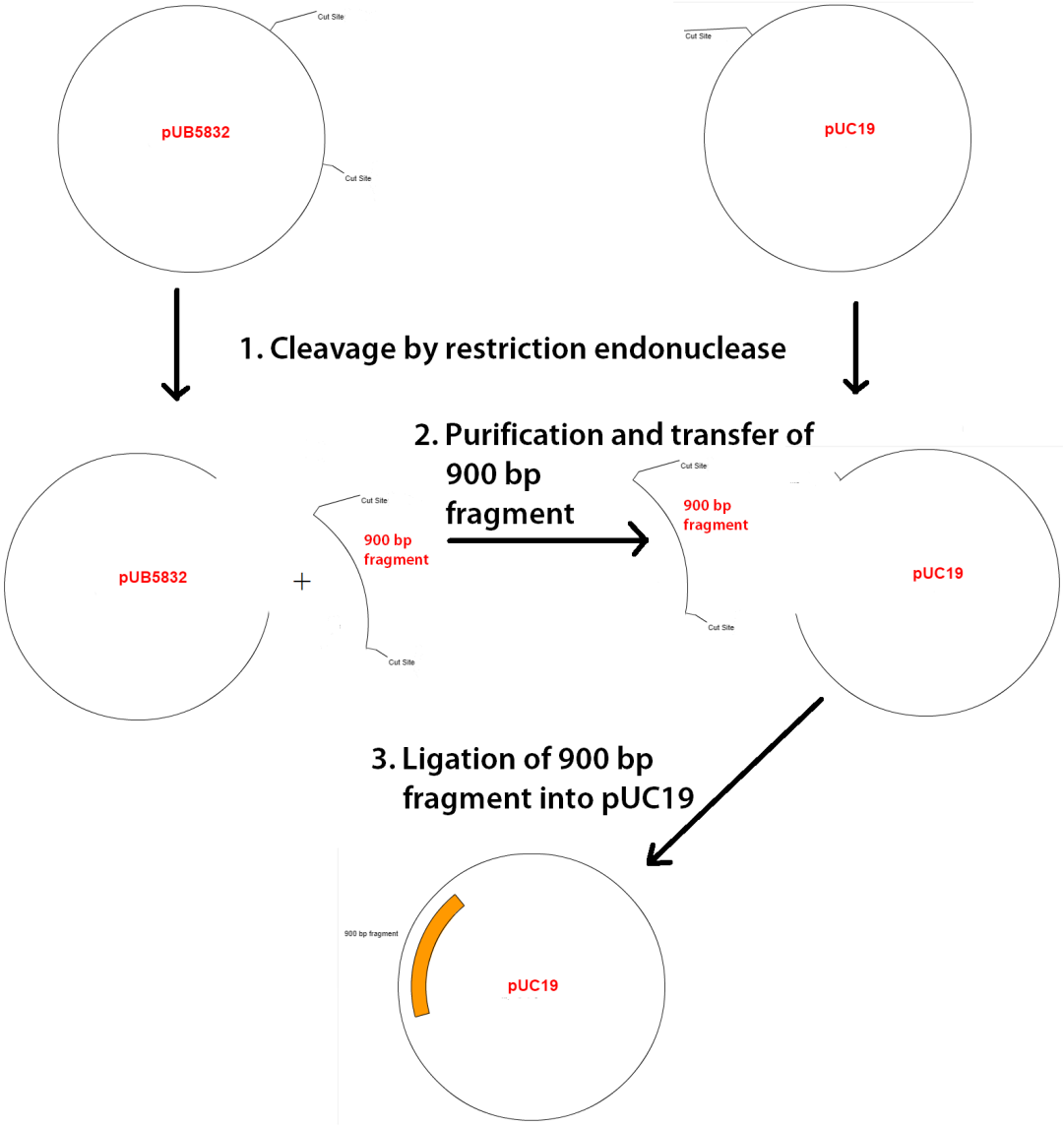

Figure 4.5: Example of a series of a sequence of events

4.6 The Derivation of Our Set of Semantic Roles

It is important to provide some detail as well on the process that led to our finalizing the choice of verb frames to be used, with their accompanying semantic roles. As was the case in determining the set of rhetorical moves, we needed to examine manually a set of sample biochemistry texts. At this point we decided to engage a domain expert in biochemistry, someone who in fact was pursuing his PhD in biochemistry. In this section we describe some of the challenges we faced. Appendix B provides more detail on some of the labelling that was done, leading to the final decisions for the frames and roles. It demonstrates the considerable effort expended in resolving this part of our knowledge representation and also serves to present a number of specific examples of how texts can be labelled with their roles.

Following our analysis in Chapter 3, we came in contact with a biochemist, a domain expert, who has expertise in the biochemistry field and is pursuing a PhD degree in biochemistry. Our communication lasted for 4 months and was fairly intense work because we knew that this person could not be our domain expert beyond that time span. We began our initial meetings to develop a common ground of what we are interested in for this domain and what type of analysis we are after. So, we explained that we aim to build a computational framework capable of analyzing and extracting relevant information from biochemistry articles. In order to build such a framework, we said that we need first to understand the text of biochemistry articles especially in the Method section. As a computational linguist, we also introduced our expert to the concept of an author's argumentation and how rhetorical moves throughout the article are being used as the blocks to describe steps of experimental procedures, hence building arguments in writing. Thus, we described our observations about the verbs used in sentences in the Method sections and how non-technical verbs used in specific senses articulated this genre, biochemistry articles. Our domain expert first started by teaching us some laboratory procedures which helped us properly interpret the writing. To illustrate, for a number of cases we had a mental image of what was going on as steps of an experimental procedure due to our interpretation of the words in the text, which was completely wrong. For example:

"Mutations were introduced into the L1 gene by using the overlap extension method of Ho et al. [60], as described previously [68]." [27]

If I were to ask you to label the above example, would you be able to tell exactly what are the verb and the theme in this sentence? Let us label the above sentence. Although the verb in this sentence is obvious ("introduced") we thought that "mutations" are being "introduced" or fed into "the L1 gene" but in reality what the sentence tells us is that "the L1 gene" is "mutated" using "the overlap extension method". This interpretation would

not be reached without the expertise in this domain. We discovered that a lot of domain knowledge is not in the text because the writing style expects experts to be reading the text. Another example that we could not interpret correctly without the help from the domain expert confirms the earlier point that not all relevant information is presented in the text. Before continuing, we provide some biochemisty knowledge (similar to what was provided to us by our domain expert) in order to interpret the example that follows.

First, plasmids, named pUB5832 and pUC19, (a plasmid is typically a small circular DNA strand in the cytoplasm of a bacterium or protozoan that are much used in the laboratory manipulation of genes) are digested (that is they are cut (or cleaved) at specific places and the pieces are removed) by two enzymes, namely *Nde*I and *Hind*III. The length of the piece removed is 900 bp (base pairs are the biochemical units in DNA). Then, the desired 900 bp piece needs to be isolated (through a procedure called gel purification). This piece is then transferred from pUC5832 to pUC19. It is inserted into the missing section of pUC19 through a process called ligation (the joining of two DNA strands) using an enzyme, T4 ligase. The modified plasmid is given a new name. These biological and laboratory procedures have been illustrated in the Figure 4.5 to help with this explanation. Multiple copies of this new plasmid will be generated (cloned).

This example is presented in text that conveys this sequence of steps in an experiment as follows:

"The over-expression plasmid for L1, pUB5832, was digested with *Nde*I and *Hind*III, and the resulting ca. 900 bp piece was gel purified and ligated using T4 ligase into pUC19, which was also digested with *Nde*I and *Hind*III, to yield the cloning plasmid pL1PUC19." [27]

Not only did we spend a lot of time being trained in understanding the text but we had to train the domain expert to understand what we were trying to accomplish (e.g., the domain expert understood our language incorrectly in many instances and we had to explain aspects of computational linguistics so that we were "on the same page"). Both directions of education finally allowed us to attack the problem of coming to agreement on the set of semantic roles that are presented in this chapter. This was a non-trivial exercise. We had many disagreements, and discussed/argued at length before coming to a consensus on each of the new semantic roles. It is important that many of these disagreements were due to not understanding the other's way of looking at the problem.

With the domain expert, we analyzed in great detail three articles (see Appendix B for the complete annotation of these three articles as well as some comments that we noted during our interaction with the domain expert), and some of the proposed semantic roles (described earlier in Section 4.1.2) emerged.

Chapter 5

Annotation

This chapter describes the steps that were conducted to complete one of the important tasks in building a knowledge representation framework (i.e., the creation of gold standard corpus). This task is used as a ground truth to validate and evaluate our framework. Such an annotated corpus is required for training of Machine Learning (ML) algorithms. This task involved human inputs and domain experts; that is a laborious[1] and time-consuming process[2]. Most of decisions in this chapter were based on the observation and analysis of the experimental procedures in the biochemistry genre. This chapter is structured as follows: Section 5.1 describes the initial analysis of the Methods section in biochemistry articles, while Section 5.3.1 explains the developed scheme for experimental events. Section 5.2 shows the creation of the biochemistry corpus and Section 5.3 discusses the process of creating the annotated corpus.

5.1 Analysis of Experimental Procedures

The initial step towards understanding the experimental procedures described in the biomedical articles for non-experts is to ask a domain expert to review the steps of that particular experimental procedure and then explain it to the non-experts. That is, an expert, PhD student in biochemistry, was involved in the early stage of analyzing the experimental procedures of biochemistry articles manually. For analysis purposes, we have created a data set consisting of 105 text files. These files include only the Method sections from biochemistry journal articles which were randomly selected from PubMed Central

[1] Over 13k Canadian dollars spent on this project.

[2] The duration of this annotation project was one year and six months.

(PMC). In the beginning, we allowed the expert to select articles from the dataset that covers a wide spectrum of topics ranging from *Identification of sites phosphorylated by the vaccinia virus B1R kinase* [24] to *The Metal Coordination of sCD39 during ATP Hydrolysis* [32]. Then, the expert was asked to explain the steps involved in the selected articles one by one in chronological order. This step involved many reiterations and demonstration by graphics or the use of educational videos that explained steps involved in experimental processes. We also have also created a list of frequent procedural verbs based on our observation and close discussion with the expert. In the following subsections, we will describe the main important observations, which are the implicit knowledge, the use of procedural verbs and the order for experimental processes in writing, based on our analysis with the domain expert.

5.1.1 Implicit Knowledge

One of the important aspects of describing experimental procedures in writing is the assumption of prior knowledge about the procedures. So, much important information cannot be inferred from the texts directly. This leads to misinterpretation and misunderstanding of what is really involved in the procedures. To overcome this issue, one can incorporate the knowledge from domain experts, through this process which is laborious and time-consuming. Alternatively, it is possible to use manuals of biochemistry procedures which catalogue standard experimental procedures thoroughly. In our case, we have involved in our study both options to extract all related information from the texts manually. For example,

Example 3 *"the resulting ca. 900 bp piece was* ***gel purified****..." [27].*

In the above example, one cannot fully understand how the resulting ca. 900 bp piece were gel purified unless knowing what is involved in the gel purification process. So, from the manual of biochemistry procedures, we found that there are main four steps involved in the gel purification. In each step, there are more than one substep(s) and the use of various instruments and materials. This implicit knowledge is needed to perform an adequate analysis of sentences (discussed further in Chapter 6).

5.1.2 General vs. Procedural Verbs

General English verbs (e.g., "wash") and specialized ones such as (e.g., "carboxymethylated") are used to describe the experimental procedures. However, these verbs require different semantic roles than are normally found in general English use. For instance,

No.	Sentence
1	The over-expression plasmid for L1, pUB5832, was digested with *Nde*I and *Hind*III, and the resulting ca. 900 bp piece was gel purified and ligated using T4 ligase into pUC19, which was also digested with *Nde*I and *Hind*III, to yield the cloning plasmid pL1PUC19.
2	Mutations were introduced into the L1 gene by using the overlap extension method of Ho et al. [60], as described previously [68].
3	The oligonucleotides used for the preparation of the mutants are shown in Table 1.1.

Table 5.1: Some sentences from the article **Biochem-3-_-77373** [27]

Example 4 *"Beads with bound proteins were* ***washed*** *six times (for 10 min under rotation at 4° C) with pulldown buffer . . . "[47].*

From the example above, the verb "wash" required various information about the process of washing the "Beads with bound proteins" such as times of washing, duration for the washing, temperature while the washing occurred, and the use of buffer to do the washing. The verb "wash" evokes certain semantic roles in writing some of them are similar to the ones normally found in lexical resources such as [137, 10] (e.g., Patient) and others are required for this domain (e.g., Protocol detail). We have discussed this aspect in more detail in Chapter 4.

We have also developed a small set of VerbNet-like frames for frequent procedural verbs in biochemistry (e.g., "biotinylated", "annealed", and "carboxymethylated") base on the manually analyzed data. We aim to extend the VerbNet project [137] by providing syntactic and semantic frames for procedural verbs.

5.1.3 Sequence of Events in Procedure-oriented Writing

A procedure is a *sequence* of *steps*. These steps can be totally ordered or partially ordered. Total ordering needs a means to represent the concept that one event precedes other event(s). As we have discussed early in Section 5.1.1 not all pieces of information are contained in the texts, so one cannot determine the order for sequences of events that occurred in the lab based on the writing only.

To illustrate, the sentences in Table 5.1 are three contiguous sentences in a biochemistry article. They discuss the idea of cutting a DNA piece of a plasmid, which is "a small circular and double-stranded DNA molecule that is distinct from a cell's chromosomal

Event 1	Event 2	Event 3
Sentence No. 1 • Patient: The over-expression plasmid for L1, pUB5832 • Predicate: digested • Instrument (catalyst): *Nde*I and *Hind*III	Sentence No. 1 • Patient: the resulting ca. 900 bp piece • Predicate: gel purified • Instrument (catalyst): Gel electrophoresis	Sentence No. 1 • Patient: pUC19 • Predicate: digested • Instrument (catalyst): NdeI and HindIII
Event 4	**Event 5**	
Sentence No. 1 • Patient: the resulting ca. 900 bp piece • Predicate: ligated • Instrument (catalyst): using T4 ligase • goal: into pUC19	Sentence No. 2 • Patient: the L1 gene • Predicate: introduced (mutated) • Instrument (reference type): using the overlap extension method of Ho et al.	Sentence No. 3 does not contain experimental events.

Table 5.2: Extracted events from two sentences in the article **Biochem-3-_-77373** [27]

DNA"[3], and ligate (attach) that piece to another plasmid to produce the desired protein. Table 5.2 shows five events from the sentences in Table 5.1. The events 1, 2, 3, and 4 are extracted from Sentence No. 1 and Sentence No. 2 has only Event 5, while there is no actual experimental event in Sentence No. 3. It rather simply refers to a table in the article's prior text.

Each event in Table 5.2 represents one complete experimental procedure. Also the actual sequence of experimental events in the lab does not necessarily follow the sequence that these events appear in the text. Another important aspect to note is that not all the essential information about experimental processes is found in the text, some information

[3]plasmid: Learn Science at Scitable (n.d.). Retrieved December 22, 2017, from https://www.nature.com/scitable/definition/plasmid-plasmids-28

can be implied. However, these implied pieces of information can be inferred from an ontology of standard biochemistry procedures, some of which we have developed.

5.2 Data Set

We have created a data set consisting of 105 text files. These files include only the Methods sections from biochemistry journal articles which were randomly selected from PubMed Central. We also applied some pre-processing methods to the data set such as part-of-speech (POS) tagging (i.e., GENIA Tagger[4]) and sentence parsing (i.e., BLLIP Parser[5]). We have used this data set for our initial text analysis that we described here in Section 5.1. Furthermore, in order to have consistency in writing standards and to ensure that our annotation for semantic roles and rhetorical moves would be generalizable to this particular genre, we decided to extend our data set to have high quality articles from the top journals in biochemistry. So, we contacted a librarian, who specialized in biochemistry, from the University of Waterloo to obtain access of the journal articles. First, we asked the librarian to find the top journals in biochemistry. The librarian showed a list of top ten journals between the years of 2013 to 2015 using Scopus[6]. These journals include only the top nine which University of Waterloo has subscriptions with:

1. **Cell**
2. **Genome Research**
3. **Molecular Cell**
4. **Molecular Biology and Evolution**
5. **Molecular Aspects of Medicine**
6. **Nature Medicine**
7. **Nature Methods**
8. **Nature Structural & Molecular Biology**
9. **Nature Chemical Biology**

[4] http://www.nactem.ac.uk/GENIA/tagger/
[5] http://bllip.cs.brown.edu/resources.shtml
[6] https://www.scopus.com/

We also asked for permission to download from these journals, so the librarian did review the subscriptions of these top journals and found that some publishers (e.g., Elsevier) require formal request directly to obtain access to a large number of articles since the University of Waterloo general subscriptions for these journals only allow limited number of articles to be downloaded. Therefore, we contacted these publishers and found that some of them allow access their repositories while others only allow limited number of articles to be downloaded with restrictions (e.g., 25 articles per hour). The resulting data set was composed of 3499 articles between the years of 2013 to 2015 from the aforementioned journals in biochemistry.

5.3 Annotation Guidelines

We created guidelines for annotating the Methods section in biochemistry articles. The guidelines include a description and the necessary background information of the task. The guidelines also include examples for each type of semantic role and their occurrence in the text. A list of questions supplements the guidelines to help annotators classify each sentence into its proper category. This task is done for semantic role labeling at the word level and rhetorical move labeling at the sentence level. We further supplemented the guidelines with a list of common co-factors and buffers that are normally used in the experimental procedures.

We hired experts in the biomedical domain to label the Methods section in all of the articles in our dataset using our annotation scheme. Due to resource limitations, only 5% of the total number of articles have been selected for annotation by two annotators. We include in Appendix A the complete annotation guidelines which were given to the annotators, the result of iterative development. We also include in Appendix A some questions about the annotation, observations, and meeting notes to illustrate the collaborative process. We will discuss the annotation procedure in the following section.

5.3.1 Annotation Scheme for Experimental Events

Based on our observations in Section 5.1, we have developed a new annotation scheme for identifying the structured representation of knowledge in a set of sentences describing the experimental procedures in the Method sections of biochemical articles. Several researchers have developed other forms of schemes (e.g., "bio-events" [157]) to extract biological information (e.g., gene regulation). However, a bio-event is different from our definition of

an experimental event. On the one hand, a bio-event is concerned with detection of biomolecular events within the biomedical literature, such as the identification of events that are related to given proteins [157]. In our case, an experimental event is concerned with processes and procedures that are used to investigate biological events. The experimental event is also concerned with the recognition of the biochemist's reasoning of standard biochemical procedures such as using certain instruments or specific biological materials. Our annotation scheme consists of two tiers of information. A *rhetorical move* is on the sentence or clause level while *semantic role* is on the word or phrase level. The following subsections describe these two tiers of information.

Annotators are allowed to select the text span for labeling units (e.g., rhetorical moves and semantic roles) with some constraints as follows:

1. For a sentence or clause to be qualified as a rhetorical move, it must include a main verb and stand on its own. For example:

   Example 5 *"Beads with bound proteins were* ***washed*** *six times (for 10 min under rotation at 4° C) with pulldown buffer ..." [47].*

2. A sentence or clause that is qualified as a rhetorical move, it should have at least one or more semantic roles. Given the previous example, one could label the sentence as follows: - "Beads with bound proteins" as a *theme* - "were washed" as a *predicate*, - "six times", "for 10 min", "under rotation", and "at 4°C" as protocol-details (repetition, time, condition, and temperature respectively).

5.3.2 Annotation for Rhetorical Moves

We have developed a set of rhetorical moves following Kanoksilapatham's [79, 80] work. We have described the development of these rhetorical moves in Chapter 3. That is, we have adapted and modified some of Kanoksilapatham's moves, as well as adding new more fine-grained moves to our annotation scheme. In combination, there are four major rhetorical moves concerned with the Methods section in biochemistry articles as can be seen in Table 5.3. An example of each move from our list is given below:

- "HEK293T cells were grown in DMEM." [172]. This is an example of the rhetorical move "Description of the method".
- "It has previously been demonstrated that roGFPs equilibrate predominantly with the glutathione redox couple through the action of endogenous glutaredoxins" [110]. This sentence should be labeled as "Background information".

Move type	Definition
Description-of-method	Concerned with sentences that describe experimental events.
Appeal-to-authority	Concerned with sentences that discuss the use of well-established methods.
Background information	Concerned with all background information for the experimental events such as "method justification, comment, or observation, exclusion of data, approval of use of human tissue" as defined by Kanoksilapatham (2003).
Source-of-materials	Concerned with the use of certain biological materials in experimental events.

Table 5.3: Rhetorical Moves in the Method Sections of Biochemistry Articles

- "Reagents were purchased from Sigma-Aldrich" [172]. This sentence is an example of the move "Source of materials".
- "These syntheses were performed as previously described" [172]. This example should be annotated as "Appeal to authority".

5.3.3 Annotation for Semantic Roles

As described earlier, our experimental event scheme was inspired by the annotation scheme for bio-events [158]. We based our experimental event scheme for verb arguments on the inventory of semantic roles in VerbNet [137] and modified and added new semantic roles to define our scheme. Our experimental event scheme includes: *Theme*, *Patient*, *Predicate*[7], *Agent*, *Location*, and *Goal*. An example of each of these semantic roles is given next. The word(s) which is (are) marked in **boldface** is (are) the word(s) to which the semantic role has been given. The complete set of semantic roles and their definitions in our experimental event scheme is presented in Table 5.4.

- "**We** also tested the liquid crystalline medium formed by cetylpyridinium chloride (CPCl) and 1-hexanol but had poor results in terms of sample stability." [30]. This is an example of the semantic role "Agent".

[7]Predicate is included here as part of our experimental event scheme but it is not a semantic role.

Semantic role	Definition
Agent	Generally a human or an animate subject.
Patient	Participants that have undergone a process.
Theme	Participants in a location or undergoing a change of location.
Goal:	
Physical	Identifies a thing toward which an action is directed or a place to which something moves.
Purpose	Identifies the stated purpose in a sentence for doing certain actions.
Factitive	A referent that results from the action or state identified by a verb.
Location	The physical place where the experiments took place.
Protocol-Detail:	
Time	Identifies the time or a duration of an experimental process.
Temperature	Identifies the temperature of an experimental process.
Condition	Identifies the condition of how an experimental process is performed.
Repetition	Identifies the number of times an experimental process is repeated.
Buffer	Identifies the buffer that was used in an experimental process.
Cofactor	Identifies the cofactor that was used in an experimental process.
Instrument:	
Change	Describes objects (or forces) that come in contact with an object and cause some change.
Measure	Describes an object or protocol that can measure another object(s).
Observe	Describes an object which can be used to observe another object(s).
Maintain	Describes an object or protocol which can be used to maintain the state of object(s).
Catalyst	Describes an object that can be used as a catalytic "facilitator" for an experimental event to occur.
Reference	Refers to a method or protocol that is being used.
Mathematical	Describes a mathematical or computational instrument

Table 5.4: Semantic Roles in the Annotation Scheme of our Experimental Event

- "**Cells** were resuspended in 100 mM Mes-Tris pH 6.0 buffer to a concentration of 7.5 D 600 nm units/ml." [110]. In this example "Cells" is labeled as the semantic role "Patient".
- "**The resulting cell lysate** was collected . . . " [110]. This is an example of "Theme".
- "All NMR data were collected at **the University of Minnesota NMR Center**." [30]. This should be labeled as the semantic role "Location".
- "**An extended structure** was first generated . . . " [30]. This is an example of the semantic role "Factitive".
- "A five-region spline of orders 2, 3, 3, 3 and 3 was used **to model the smoothly decaying post-edge region**." [30]. This should be labeled as "Goal:Purpose".
- "The ca. 900 bp PCR products ligated **into PUC19** . . . " [27]. In this example, "into PUC19" should be labeled as "Goal:Physical".

Working with a biochemist as described in Chapter 4, we have extended the VerbNet definition of the semantic role *Instrument* from simply "an object or force that comes in contact with an object and causes some change in them" [137] to include a variety of subcategories corresponding to various types of biological and man-made instruments used in a biochemistry laboratory. We have also added *Protocol detail* as a set of semantic roles that identify certain types of information about experimental processes such as time and temperature. These subcategories include:

- Instruments used to *change* the state of an object. For example:

  Example 6 *"Beads with bound proteins were washed six times (for 10 min under rotation at 4° C) with* ***pulldown buffer*** *..." [47].*

  In this example, the pulldown buffer was used to wash (change the state of) the Beads with bound proteins. In this instance, the phrase "pulldown buffer" should be labeled as **instrument (change)**.

- Instruments used to *maintain* the state of an object. For example:

  Example 7 *"Once the samples were in EPR tubes, they were immediately frozen in liquid nitrogen, and stored in* ***liquid nitrogen*** *before using." [32].*

In this example, the liquid nitrogen was used to store (maintain the condition of) the samples which were in the EPR tubes. In this case, the phrase "liquid nitrogen" should be labeled as **instrument (maintain)**.

- Instruments used to *observe* an object. For example:

  Example 8 *"The mitochondria was observed by* ***spinning disk confocal microscopy****."*

  The spinning disk confocal microscopy is used to observe the mitochondria. We should label the phrase "spinning disk confocal microscopy" as **instrument (observe)**.

- Instruments used as a *catalyst* in experimental processes to occur. For example:

  Example 9 *"The ca. 900 bp PCR products were digested with* ***NdeI and HindIII*** *and ligated into pUC19." [27].*

  In this example, the NdeI and HindIII are enzymes used to facilitate the digestion (cutting) of the ca.(approximately) 900 bp PCR products. In this instance, the phrase "NdeI and HindIII" should be labeled as **instrument (catalyst)**.

- Instrument used to *measure* an object. For example:

  Example 10 *"Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by* ***autoradiography****." [47].*

  In this example, the autoradiography was used to analyze (measure) the proteins. In this example, the word "autoradiography" should be labeled as **instrument (measure)**.

- It could be used to describe a *mathematical or computational instrument* (e.g., simulation, algorithm, equation, and the use of software). For example:

  Example 11 *"Simulations of these EPR spectra were accomplished with* ***the computer program QPOWA [30,31])****." [32].*

  The computer program QPOWA was used here as computational instrument to perform simulations of the mentioned above EPR spectra. So, the phrase "the computer program QPOWA [30,31]" should be labeled as **instrument (computational instrument)**.

- Finally it could be used as a *reference* for method or protocol that being used. For example:

  Example 12 *"The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase were* ***as previously described [11].*** *" [24]*

  The phrase "as previously described [11]" is to indicate that the authors referring to other method that they used in their current experimental process. We should label the phrase "as previously described [11]" as **instrument (reference)**. These subcategories of the semantic role (instrument) are not exclusive to the mentioned types above. However, based on our full-text analysis, these instrument types are most comprehensive ones. We will add or update these subcategories if we encountered a new type (usage) of instrument.

We have also proposed a new semantic role *Protocol detail* that identifies certain types of information about experimental processes which include:

- **Time** or the duration of a process [137]. For example:

  Example 13 *"Beads with bound proteins were washed six times (****for 10 min*** *under rotation at 4° C) with pulldown buffer . . . " [47].*

- **Temperature** of an experimental process. For example:

  Example 14 *"Beads with bound proteins were washed six times (for 10 min under rotation* ***at 4° C****) with pulldown buffer . . . " [47].*

- **Condition** or manner of which an experimental process was carried out. For example:

  Example 15 *"Beads with bound proteins were washed six times (for 10 min* ***under rotation*** *at 4° C) with pulldown buffer . . . " [47].*

- **Buffer** which is "a solution containing either a weak acid and a conjugate base or a weak base and a conjugate acid, used to stabilize the pH of a liquid upon dilution."[8] For example:

[8]Buffer: Biology-Online Dictionary (n.d.). Retrieved September 23, 2017 from http://www.biologyonline.org/dictionary/Buffer

Example 16 *"For phosphorylation, three identical reactions contained H5R protein (70 pmol), B1R protein kinase (90 µl),* ***Tris-HCl, pH 7.4 (20 mM)****, magnesium chloride (5 mM), ATP (50 µM), [γ-32P] ATP (50 µCi) and dithiothreitol (2 mM) in a total volume of 500 µl." [24].*

- **Cofactor** is defined as "inorganic substances that are required for, or increase the rate of, catalysis."[9] For example:

  Example 17 *"For phosphorylation, three identical reactions contained H5R protein (70 pmol), B1R protein kinase (90 µl), Tris-HCl, pH 7.4 (20 mM),* ***magnesium chloride (5 mM), ATP (50 µM), [γ-32P] ATP (50 µCi) and dithiothreitol (2 mM)*** *in a total volume of 500 µl." [24].*

- **Repetition** of a step in experimental processes. For example:

  Example 18 *"Beads with bound proteins were washed* ***six times*** *(for 10 min under rotation at 4° C) with pulldown buffer . . . " [47].*

5.3.4 Human Input and Annotation Procedures

We advertised the annotation study to the faculty of Science in the University of Waterloo. We were looking for graduate and undergraduate, who are in the 3rd or 4th year of their studies, students. We interviewed each candidate and asked for their credentials. So, we hired ten annotators with a variety of backgrounds (Biochemistry, Bioinformatics, Biology) and different academic levels ranging from Bachelor to PhD degree. The annotators engaged in various training sessions that were led by the author. We provided different resources that can help and support the annotators in this project. These resources include frequent meetings, the annotation guidelines, a list of questions and answers about the annotation (see Appendix A), our biochemistry expert (a PhD student working with us), and the use of web-based software called Slack[10] which allows annotators to post questions, comments, or illustrate an example from the data set. We have also created a demo video[11] that shows annotators step by step how to use the GATE tool[12] and how to use the schema (the list of xml schema that are used for annotation is in Appendix C) to

[9] coenzymes and cofactors. (n.d.). Retrieved September 23, 2017, from http://academic.brooklyn.cuny.edu/biology/bio4fv/page/coenzy$_$.htm

[10] https://slack.com/

[11] At https://uwaterloo.ca/scholar/mallihee/links/gate-annotation-demo-and-annotation-guidelines

[12] https://gate.ac.uk/

label texts. This was a very important decision to help to address the experts' knowledge gap in computational linguistics, and to facilitate comprehension of the interface that was provided. Then, we set up several training sessions for the expectations of this study. Essentially, each annotator is asked to read the guidelines and if at any point she/he has a question or needs clarification, we can illustrate by providing more examples. We set up a meeting with the annotators either by Skype or in person to answer their questions. In fact, the guidelines have been revised and updated several times to reflect the annotators' feedback.

Annotators are asked to download and use the GATE tool as an interface which gives them access to our developed schema[13] for the semantic roles and rhetorical moves. Each article is labeled by two annotators. The labeling is done on a verb basis rather than a full-sentence basis. In other words, each sentence with more than one verb is divided into smaller text spans (Annotation Units (AUs)), which are composed of a verb and the text containing its semantic roles. The annotators identify the verb in that AU and label all associated semantic roles for that verb within that AU. The annotators decide which constituent is a semantic role. Then, annotators label the entire AU with appropriate rhetorical moves. Each annotation is stored in an XML file. Figure 5.1 shows an example of some sentences annotated for both rhetorical moves and semantic roles.

5.4 Inter-annotator Agreement

5.4.1 Identification of Semantic Roles

We measured the inter-annotator agreement for semantic role labeling between the two annotations of the same article using the κ-score [33]. To have a matching label, both the semantic role category and the text span must be the same. Then, we measured the κ-score after the adjudication step which was done by us. The adjudication step's main goal is to resolve any disagreement in annotations [115]. The adjudication step includes correcting mislabeled spans into the proper category and correcting the extent of labeled spans for specific semantic roles[14]. We have also measured the kappa score for different configurations of the data set as shown in Table 5.5. "Original annotation" is the annotation that was provided by the annotators. "Theme combined with patient and all instrument roles combined" indicates theme and patient were combined as one role and

[13]The developed schema is supplemented in Appendix C

[14]Whereas in the work of Palmer et al. [115], adjudication was performed by small team of highly trained linguists, for our domain we need to engage the annotators to resolve the adjudication

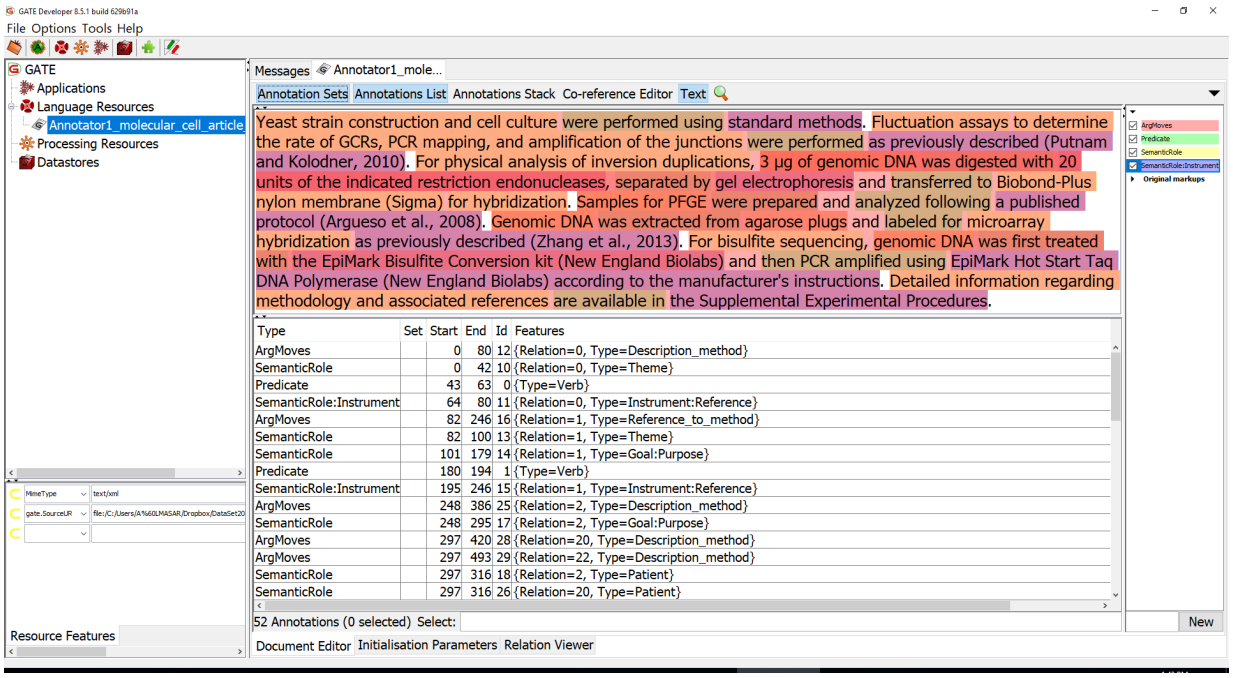

Figure 5.1: Snippet extracted from one article [41] of our annotated dataset showing the labelling of the rhetorical moves and semantic roles using the GATE tool.

all instrument subcategories were considered as one. "Protocol detail combined" indicates that in addition to the previous merging of semantic roles, all protocol detail subcategories were combined as one role. "Adjudicated" means that the disagreements in the original annotations were resolved and any missing semantic roles were added. All of the κ-scores in Table 5.5 are rated substantial [89, 106]. The results are very promising.

5.4.2 Identification of Rhetorical Moves

We also measured the inter-annotator agreement for rhetorical move identification between the two annotations of each article using the κ-score. Here again, the rhetorical move and text span must be the same to be considered a match. As seen in Table 5.6, we have measured the kappa-score for two configurations. "Original" is the annotation provided by the annotators, while "Adjudicated" means that the disagreements in the original annotations were resolved. The result, shown in Table 5.6, shows a moderate to almost perfect agreement [89, 106]. We have calculated the confusion matrix for the original annotation of rhetorical moves in Figure 5.2. During our adjudication step, we noticed some commonly

Configuration	Kappa score
Original annotation	61.3%
Theme combined with patient and all instrument roles combined	68.9%
Protocol details combined	71.6%
Adjudicated	93.6%

Table 5.5: Inter annotator agreement κ-score for semantic role labeling

Configuration	Kappa score
Original	42.0%
Adjudicated	98.2%

Table 5.6: Inter annotator agreement κ-score for rhetorical move identification

mislabeled instances by some annotators. For example:

Example 19 *"The hierarchical cluster analyses were performed in MATLAB (Release 2012a), and the bar graphs were produced in Microsoft Excel 2010." [39].*

This sentence should be labeled "Description-of-method" since it clearly describes steps of the authors' method, i.e., using tools to perform analyses and produce graphs. However, one annotator mislabeled it as "Appeal-to-authority".

Example 20 *"Constructs comprising new opsin sequences cloned in pMT4 were transiently transfected into Neuro-2a cells with GeneJuice reagent (Novagen), according to the manufacturer's instructions (for further information, see Supplemental Material)." [39].*

This sentence was labeled incorrectly as "Description-of-method" whereas it should be labeled as "Appeal-to-authority" since it refers to an "established" method.

5.5 Remarks

In this chapter, we have presented examples of semantic roles that we have suggested to be necessary for this scientific domain and which are used in our annotation scheme. This Experimental Event Scheme, which is based on the proposed semantic roles, is the first step towards developing an automated rhetorical move analysis. We have also presented the most common rhetorical moves based on our manual analysis and observations of biochemistry procedures. We also have described our annotation study along with the dataset used. Ultimately, we aim to develop a framework to analyze argumentation structure in biochemistry procedures using the rhetorical moves. We have concluded that our annotation guidelines need to be updated to better aid our annotators to properly select the right rhetorical move for each candidate AU. We note that while there is substantial agreement among annotators in our results with respect to semantic roles, the agreement regarding rhetorical moves is more modest. One reason why this might be the case is the fact that the annotated dataset to date is relatively small and annotators might actually have more inherent insight into recognizing the differences between rhetorical moves. Since these moves have spans which range from clauses to full sentences, whereas semantic roles are confined to at most a few words, the guidelines for annotation that were developed focused more on this simpler case. We anticipate expanding these guidelines in order to improve inter-annotator agreement regarding rhetorical moves in the future.

As future work, in parallel with annotating the complete data set, we will develop a computational model to label the rhetorical moves for this domain. As well, from our experience with annotating the biochemistry articles with our experts, we recognized that not all of the information needed to interpret the move structure is available in the text. What is needed is an ontology that captures the knowledge that a working biochemist would have regarding biochemistry experimental procedures, especially the sequence of events that are normally undertaken in these procedures. We have begun building such an ontology and future development will involve some automation. This is described in the very next chapter, Chapter 6.

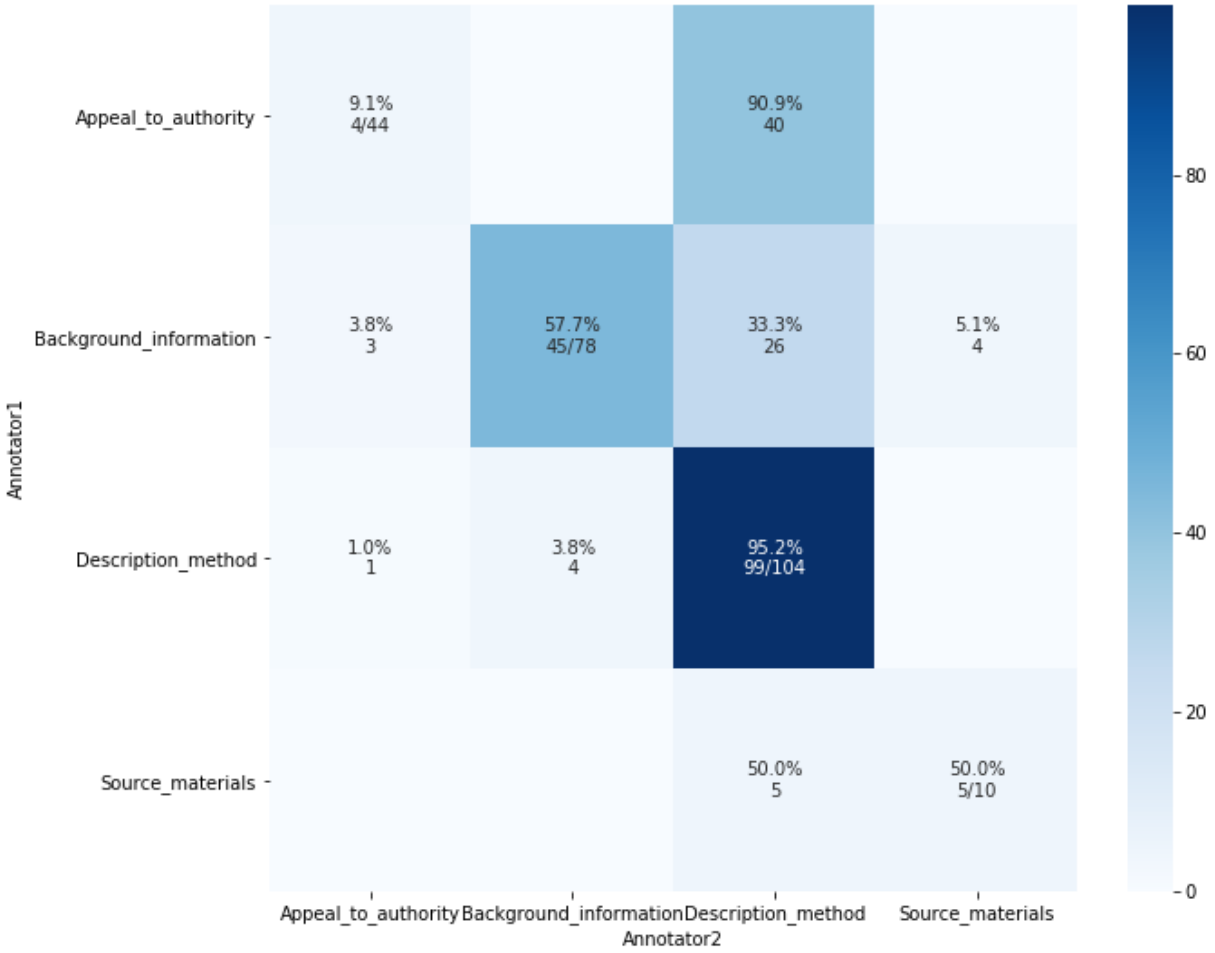

Figure 5.2: Confusion matrix for rhetorical move identification.

Chapter 6

Ontology

This chapter describes the development of our ontology for analyzing experimental procedures in the biochemistry articles. Ontologies must provide the entities, concepts, and relations required by the domain being represented. The ontology language being used is OWL-DL. OWL-DL was adopted due to its well-balanced flexibility among expressiveness (e.g., class description, cardinality restriction, etc.), completeness, and decidability [105]. These procedures are composed of procedure steps which can be represented as sequences. Sequences are composed of totally ordered, partially ordered, and alternative subsequences. Subsequences can be represented with two relations, *directlyFollows* and *directlyPrecedes* that are used to represent sequences. Alternative subsequences can be generated by composing a *oneOf* function in OWL-DL, referred to it as *optionalStepOf* in this work, which is a simple generalization of *exclusiveOR*. Alkaline Agarose Gel Electrophoresis, a biochemistry procedure, is described and examples of these subsequences are provided.

6.1 Background Information

Ontologies provide entities (known as individuals in some ontological languages) and concepts, and relations among those entities and concepts. Ontologies must provide relations that are required by the domain being represented. Our interest is centered on the biochemistry domain, the experimental methodology aspect, in particular.

A number of biologically oriented ontologies have been created; one of the best known is the Gene Ontology (GO) [7]. Others have been developed for a variety of other purposes. They are discussed in detail in the next section. Most of these ontologies describe a set

of concepts and categories in the biological domain that shows their properties and the relations between them.

The type of domain that we are attempting to represent consists of *procedures*, experimental procedures, in particular. Procedures are *sequences* of *procedure steps* (simply, *steps*, henceforth). Some ontologies provide descriptions of steps [144]. To the best of our knowledge no current biologically oriented ontology represents sequences of steps. An important aspect of the steps in a procedure is that they immediately follow one another. 'Directly follows' (and 'directly precedes') is an intransitive relation (i.e., if B directly follows A, and if C directly follows B, then C does not directly follow A). Transitive relations are the norm in the current biologically oriented ontologies (e.g., the omnipresent 'subclass' relation; 'proper part of', 'precedes' and 'is causally related to' ([44], Figures 6 and 9)).

Procedures can contain sequences of steps that are totally ordered (i.e., the steps must be done one after the other in the sequence specified), steps that can be partially ordered (i.e., subsequences of steps that can be done in any order), and alternative subsequences of steps (i.e., only one of the alternatives is done). In addition to the intransitive relations 'directly follows' and 'directly precedes' our contribution also includes these three types of sequence orderings.

Descriptions of experimental procedures exist in scientific writing. The scientific domain of interest to us is biochemistry. An important type of information contained in the Method section of biochemistry articles are references to standard biochemistry experiment procedures. These protocols, which typically involve several steps, are described in detail in manuals of standard biochemistry experiment procedures [22, 133]. In this thesis, we propose a biochemistry procedure-oriented ontology that explicitly identifies all of the steps of an experimental procedure and provides the relations between the steps of an experimental procedure. A case study investigates one experimental procedure, Alkaline Agarose Gel Electrophoresis, that exists in the manual of standard biochemistry experimental procedures. Appendix D displays this case in detail.

6.2 Related Work

Developing ontologies has become increasingly crucial in the biomedical domain in general [131]. Several ontologies have been developed in recent years such as the Gene Ontology [7], the Ontology for Chemical Entities of Biological Interest (ChEBI) [40], the Ontology for Biomedical Investigations (OBI) [12], and the Foundational Model of Anatomy (FMA) [131]. Mainly, the goal of these ontologies is to provide definitive controlled terminologies that describe entities in the biomedical genre.

The main aspect of Gene Ontology (GO) is to provide information that describes gene products using precisely defined vocabulary [7]. GO intially used three model organism databases including FlyBase [54], Mouse Genome Informatics [17, 129], and the saccharomyces Genome Database [11]. Recently, the number of model organism databases has increased dramatically [55].

The Chemical Entities of Biological Interest ontology (ChEBI) is a lexicon of molecular entities concerned with small molecules [40]. To create ChEBI, data from several resources (e.g., IntEnz [53], KEGG COMPOUND [77], and the Chemical Ontology) were used. ChEBI used various relations to describe the relationships between ontology entities. These relations include relations required by ChEBI (e.g., 'is conjugate acid of', and 'is tautomer of') as well as relations which are defined by the Relations Ontology[1] (e.g., 'is a' and 'is part of').

The Ontology for Biomedical Investigations (OBI), http://purl.obolibrary.org/obo/obi, [12], a resource for annotating biomedical investigations, provides standard tools to represent study design, protocols and instrumentation used, the data generated and the types of analysis performed on the data. Several ontologies [37, 23, 179, 144, 44] are based on the OBI ontology. These ontologies are closest to our interest in biochemistry procedures.

A work predating the above list, [145], proposes EXPO, an ontology of scientific experiments, in general. It remains a descriptive ontology, providing a detailed description of various aspects of scientific experiments and how they are related.

Descriptions of experimental processes are provided by OBI, and three real-world applications are discussed in [23]. Some of the relations in these applications (e.g., inputs, outputs, etc.) come very close to our purpose here. The beta cell genomics application ontology (BCGO) [179] also uses OBI, but it tends to be a more descriptive ontology than some of the others that use OBI, but some of the relations in RO, the relation ontology [142], that are used (e.g., produces, translate_to) do have an ordering sense.

The two ontologies that are most similar to the work described below are EXACT [144] and the Semanticscience Integrated Ontology [44]. Both are motivated by a need to describe scientific protocols and experiments. Where they differ from what we are proposing is that they describe *sets* of actions in scientific protocols and experiments, whereas we are proposing to represent *sequences* of actions, or steps in a procedure, if you like. Relations that describe orderings of actions (e.g., 'precedes' [44]) are not applicable to sequences since these relations are transitive.

[1] http://www.obofoundry.org/ontology/ro.html

The Molecular Methods Database (MolMeth) is a database which contains scientific protocol ontologies that conform to a set of laboratory protocol standards [86].

Other ontologies describe general concepts that are useful to a biochemistry procedure-oriented ontology include: Ontologies consist of process such as [94] and [135], ontology for units of measure [127], classification of scenarios and plans (CLASP) [42], and materials ontology [8]. Foundational theories such as process calculus and regular grammar are essential for the formalization of procedure-oriented ontologies.

6.3 Procedure-oriented Ontology

We propose a framework for procedure-oriented ontologies that explicitly identifies all steps of an experimental procedure and provides a set of relations to describe the relationships between the steps of an experimental procedure. The novelty of this approach is to allow creating a sequence of events (or steps in a procedure) using the ontological concept of "something occurs before". To accomplish this we need to have an ontological concept of "sequence". This is very significant concept because one cannot simply call a sequence of events "a sequence" unless these events happen step by step in some sort of ordering.

This approach will be used to provide the necessary information about the experimental procedures for Knowledge Base systems with the required knowledge about experimental processes. There are manuals of standard procedures in biochemistry [22, 133] which in turn will help in building ontologies.

6.3.1 Classes and Properties

The proposed ontology framework consists of three core classes: Step, State, and Action.

Step

The Step class (see Figure 6.1) represents each step within a procedure. Orderings of each step can be described by object properties such as 'precedes', 'follows', 'parallel', all being transitive. The properties 'precedes' and 'follows', inverses of each other, indicate the chronological order of the steps. The property 'parallel' is symmetrical which indicates steps can happen simultaneously. Intransitive properties 'directlyPrecedes' and 'directly-Follows' are also used to describe the ordering of steps. They are subproperties of 'precedes' and 'follows' respectively. Similar to 'precedes' and 'follows', they are also inverses of each

other. Therefore, by stating step1.1 'directlyPrecedes' step1.2 and step1.2 'directlyPrecedes' step1.3, a reasoner will automatically infer that step1.1 'precedes' step1.2 as well as step1.3. Also, step1.3 'directlyFollows' step1.2 but only 'follows' step1.1, both being inferable by a reasoner. For cleanliness, we indicate only the 'precedes' relation in the figures presented in Figure 6.1.

The structure of the procedure is outlined by the properties 'subStepOf' and 'optionalStepOf' in which both domain and range of the properties are Step. 'subStepOf' indicates that the step(s) must be completed for the completion of the parent step, e.g., the triples (step1.1, subStepOf, step1) and (step1.2, subStepOf, step1) state that step1.1 and step1.2 must be completed in order to consider step1 to be completed. Conversely, 'optionalStepOf' indicates that one of the steps (not both) must be completed in order to complete the parent step, e.g., (step1.1a, optionalStepOf, step1.1) and (step1.1b, optionalStepOf, step1.1) state that one and only one of step1.1a or step1.1b needs to be completed to complete step1.1.

Figure 6.1 illustrates a scenario in which all individuals are Step instances. Also, step1 is parallel to step2 while step1.1 must complete before step1.2. Note, there are no ordering relations between step1.1.1 and step1.1.2 since they are optional steps of step1.1.

State and Action

The class Step with corresponding properties outlines the structure of a procedure. The actual process in each step is represented as states and their associated actions. Each step involves a transition from state to state via a single or a series of actions, represented by the classes State and Action (see Figure 6.2). State is connected to Step via the property 'hasState' and has three subclasses, InitialState, MidState, and FinalState which are connected via properties such as 'precedes' and 'follows'. InitialState can only precede a state while FinalState can only follow another state. Triples (StateX, precedes, StateY) imply (StateY, follows, StateX), and vice versa, since 'follows' is an inverse property of 'precedes'. Figures 6.1 and 6.2 omit 'follows' to keep the figures clean. MidState can be connected to another state with both 'precedes' and 'follows' properties. Note that a step has at most one instance of InitialState or FinalState but may have multiple instances of MidState. For example, an instance of Step, step1, may involve two instances of State, i.e., step1_state1 and step1_state2, represented by the following triples: (step1, hasState, step1_state1), (step1, hasState, step1_state2), (step1_state1, precedes, step1_state2).

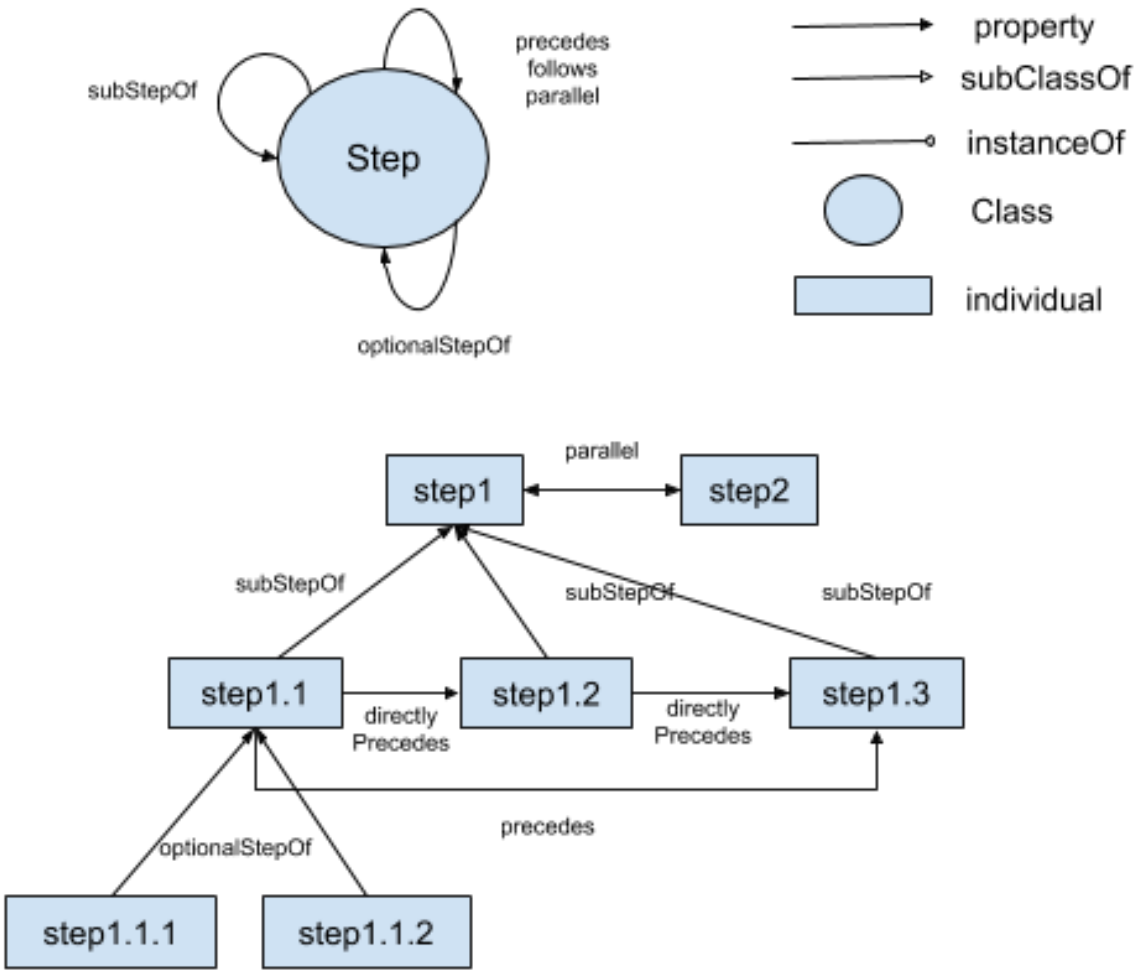

Figure 6.1: Step class and example instances

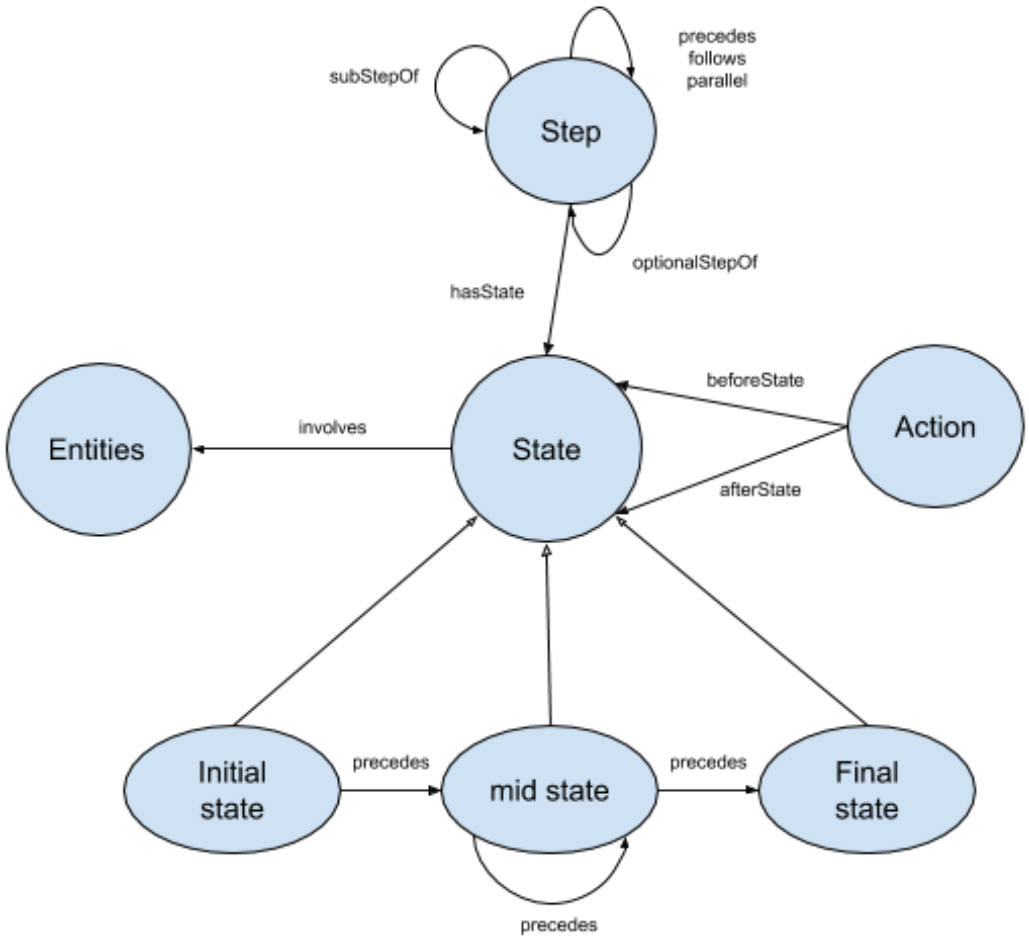

Figure 6.2: State and Action classes

Biochemistry Domain Knowledge

States are connected to the Action class via 'beforeState' and 'afterState', representing the states before and after an action, respectively. The State class is also connected to the Entities class (see Figure 6.3) via the property 'involves' which can be expanded to describe instruments, materials, and devices involved in a specific state. Thus, domain knowledge of biochemistry can be described by extending the Entities class. For demonstration purposes, we have only included selected general concepts related to experimental procedures described in the Case Study. Instrument includes Container and Device where Container 'contains' Material which is a class for Chemical and Non-Chemical materials used in biochemistry experiment procedures. Compound materials and assembled instruments are represented using the property 'consistsOf'. Instrument and Material can be connected to the class Measure which is a combination of numerical values and Unit_of_Measure, e.g., '10m' is a measure where the value is 10 with a unit of measure of 'meter' [127]. The Measure class was extended with subclasses to represent absolute measures (e.g., 10m), range values (e.g., 5m-10m), and ratio (e.g., 1/2).

6.3.2 Relations

We first need to examine the types of features that an experimental procedure needs for its definition. A procedure is a *sequence* of *steps*. These steps can be totally ordered or partially ordered. Total ordering needs a means to represent the concept that one event precedes another event and this relation needs to be transitive. Because a procedure is a sequence of steps, there needs to be a means to represent the relation that one step immediately follows another step and this relation needs to be intransitive. These relations have been defined for OWL and are available from `http://www.ontologydesignpatterns.org/cp/owl/sequence.owl`. Partial ordering is accomplished simply by allowing more than one step to follow or to precede another step. Finally, we would like to be able to represent a subsequence of steps and the choice of a subsequence from one or more possible subsequences. This 'optionalStepOf' relation would need to be crafted depending on how many choices are available. If two choices, this relation is simply equivalent to exclusive or otherwise it is simply a generalization of the exclusive or. We have developed the concept of "procedure" based on these underlying relations.

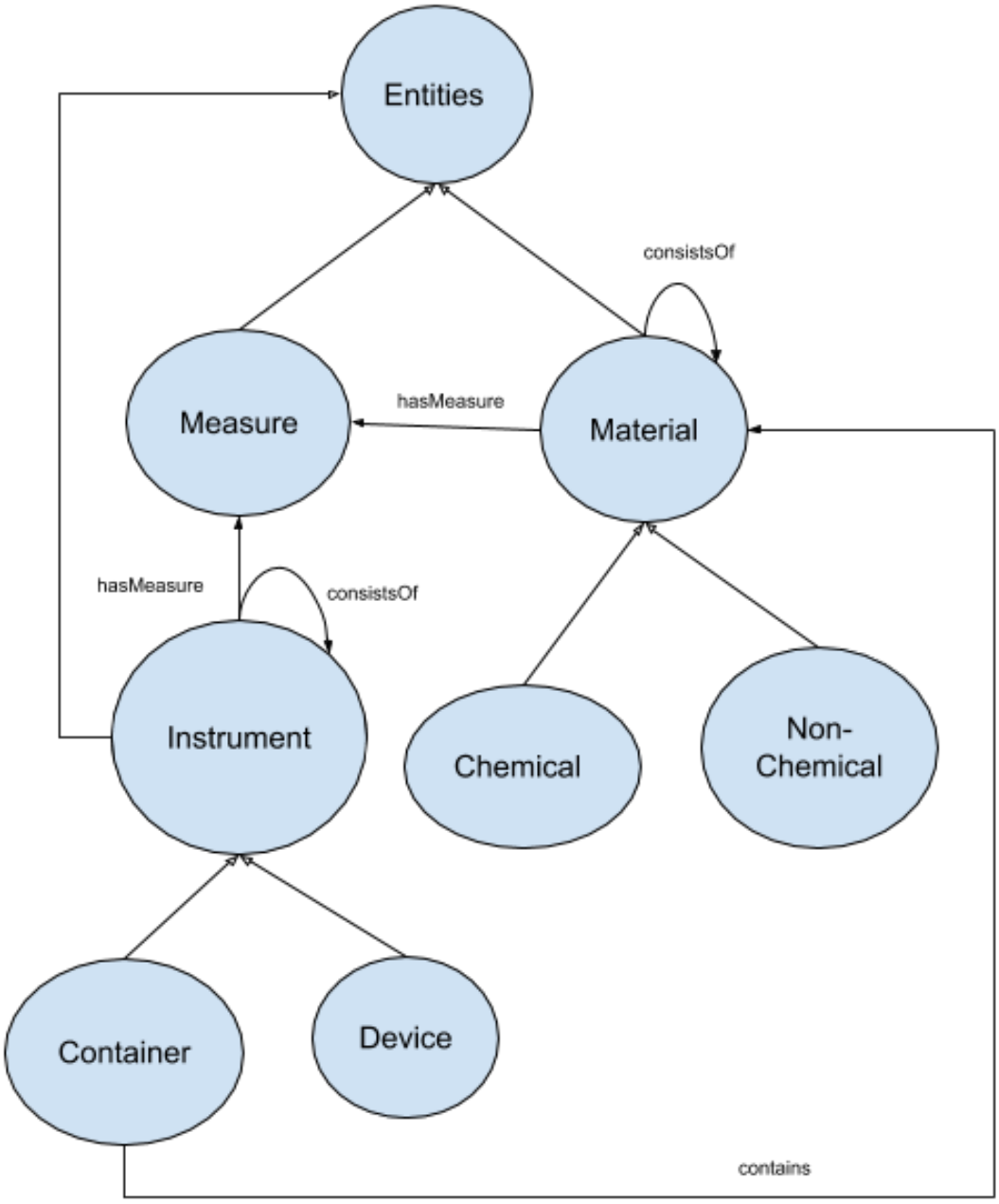

Figure 6.3: Demonstration of Entities class

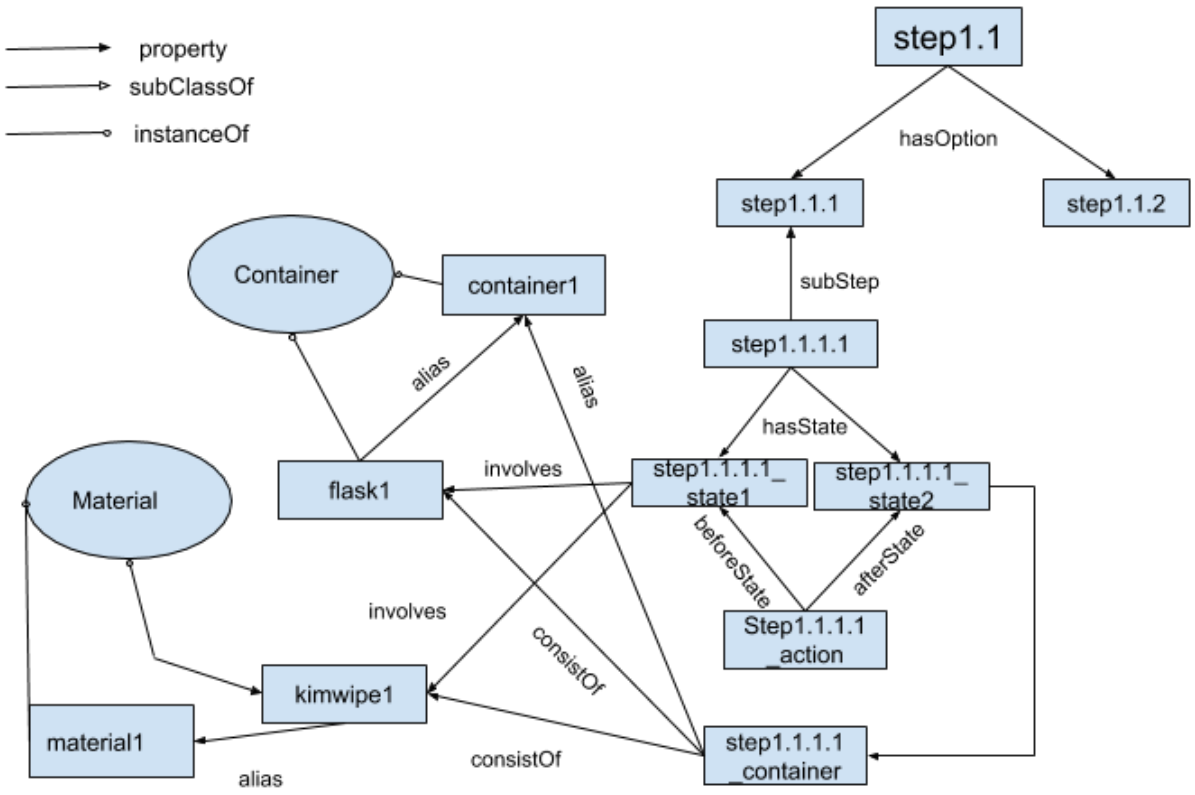

Figure 6.4: An example of alternative sub-sequences in steps for preparing the Agarose solution

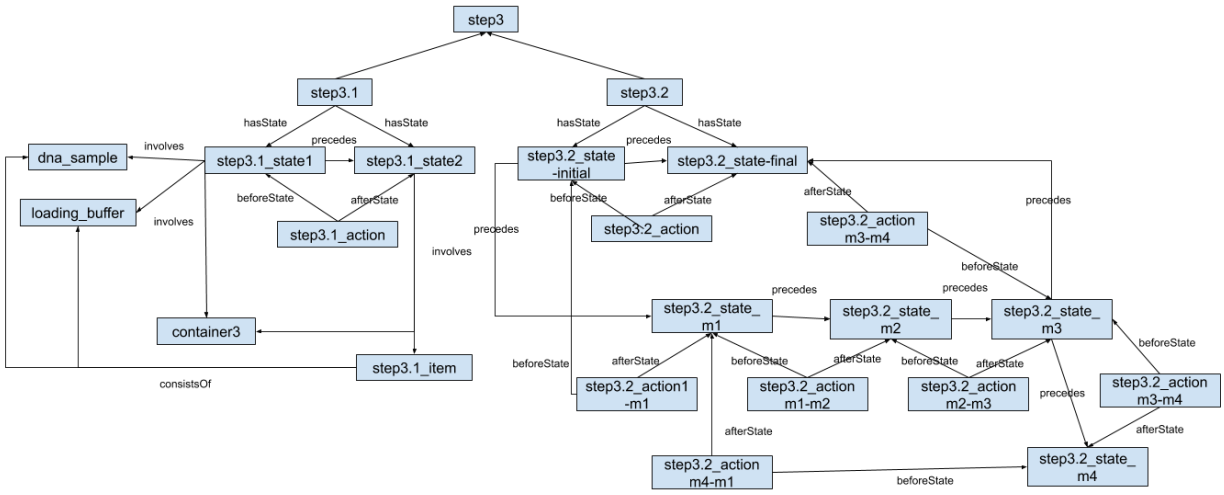

Figure 6.5: Instances related to Step3 which involves initiating the electrophoresis

6.4 Case Study

We have designed a procedure-oriented ontology for Alkaline Agarose Gel Electrophoresis [133] using the set of relations described in Section 6.3. Our ultimate goal is that we can use the ontology to fill missing information in the verb frame of a sentence. Our motivation is analyzing the text in the Method section of biochemistry articles. Since the Method section in biochemistry articles is describing experimental procedures, these procedures use some steps that are not explicitly mentioned in the text because the article is intended for readers who have prior knowledge of the field. Thus, without knowing this implicit information, one cannot fully understand all the steps of experimental procedures. For example, in order to understand fully the sentence fragment, "the resulting ca. 900 bp piece was gel purified and ligated using T4 ligase into pUC19" [27], one needs to access the information involved in gel purification and ligation. Thus, we have moved to build an ontology that satisfies this requirement.

Appendix D shows the complete steps of Alkaline Agarose Gel Electrophoresis that are involved in preparing both the agarose solution and the DNA samples. Figure 6.4 describes step1.1, the preparation of the agarose solution. Basically, step1.1 "adding the appropriate amount of powdered agarose to a measured quantity of H2O" has two options either: step1.1.1 "an Erlenmeyer flask" 'exclusiveOR' step1.1.2 "a glass bottle". So we have a relation that conveys the choice of using one container or another. So, there is a choice of two sequences of steps: *If* step1.1.1 "an Erlenmeyer flask" is selected *then* 'directlyFollows'

step1.1.1.1 “loosely plug the neck of the Erlenmeyer flask with Kimwipes” which involves both initial and final states, action and container as seen in Figure 6.4; *else if* step1.1.2 “a glass bottle” is selected *then* ‘directlyFollows’ step1.1.2.1 “make sure that the cap is loose”. In future steps of the ontology, the instance Container1 appropriately refers to the instances of either Erlenmeyer flask or the glass bottle and material1 refers to the instances of kimwipes or glass bottle cap. The two main steps (step1, and step2) shown in Figure 6.1 are meant to be partially ordered, that is, they can be performed in any order (i.e., step1 then step2 or vice versa). In addition, each one of these main steps consists of several steps (mini-steps or sub-steps).

Subject	Property	Object	Description
step3.2_state_initial	rdf:type	InitialState	
	involves	electrophoresis	
	involves	electrophoresis_measure	
	precedes	step3.2_state_m1	
step3.2_action_initial_m1	rdf:type	TurnOn	TurnOn is a subclass of Action
	beforeState	step3.2_state_initial	
	afterState	step3.2_state_m1	
step3.2_state_m1	rdf:type	MidState	
	involves	electrophoresis	
	involves	electrophoresis_measure	measure for the migration of bromocresol green
	involves	bg_migrate_measure	
	involves	bromocresol_green	
	involves	gel	
	precedes	step3.2_state_m2	
step3.2_action_m1_m2	rdf:type	DoNothing	DoNothing is a subclass of Action
	beforeState	step3.2_state_m1	
	afterState	step3.2_state_m2	
step3.2_state_m2	rdf:type	MidState	measure for the migration of bromocresol green; measure of current gel length that the bromocresol green has migrated to
	involves	bg_migrate_measure	
	involves	bromocresol_green	
	involves	gel	
	involves	gel_length_portion	
	precedes	step3.2_state_m3	
step3.2_action_m2_m3	rdf:type	TurnOff	
	beforeState	step3.2_state_m2	
	afterState	step3.2_state_m3	

Continued on next page

Continued from previous page

Subject	Property	Object	Description
step3.2_state_m3	rdf:type	MidState	
	involves	electrophoresis	measure of current length of gel that the bromocresol green has migrated to, less than 2/3
	involves	electrophoresis_measure	
	involves	gel_length_portion	
	precedes	step3.2_state_m4	
	precedes	step3.2_state_final	
step3.2_action_m3_m4	rdf:type	Action	Put glass plate on gel
	beforeState	step3.2_state_m3	
	afterState	step3.2_state_m4	
step3.2_state_m4	rdf:type	MidState	
	involves	gel	
	involves	gel_length_portion	
	involves	glass_plate	
step3.2_action_m4_m1	rdf:type	TurnOn	
	beforeState	step3.2_state_m4	
	afterState	step3.2_state_m1	
step3.2_action_m3_final	rdf:type	Action	Put glass plate on gel
	beforeState	step3.2_state_m3	
	afterState	step3.2_state_final	
step3.2_state_final	rdf:type	FinalState	measure of current length of gel that bromocresol green has migrated to, equal to or more than 2/3
	involves	electrophoresis	
	involves	electrophoresis_measure	
	involves	gel_length_portion2	
	involves	bromocresol_green	
	involves	gel	

Table 6.1: Description of the entities involved in Step3.2

As one can see, Figure 6.4 shows a total ordered sequence. Another example, shown in Figure 6.5, describes the instances of step3, step3.1 and step3.2 that are concerned with initiating the electrophoresis. Step3.1 is straightforward. Since step3.2 involves a condition to ensure the gel reaches a certain length, this step requires several MidStates in addition to both the initial and finial states as is shown in Table 6.1. All entities for step3.2 are described in Table 6.1. Note that Step3.2 consists of a number of MidStates which represents waiting until the desired amount of migration has been reached (i.e., 2/3 of gel length). The instance `step3.2_state_initial` and `step3.2_state_final` are

instances of **InitialState** and **FinalState**, respectively. The instances of **MidStates** are `step3.2_state_m1` to `step3.2_state_m4`, each representing a middle state described below:

- `step3.2_state_m1`: Electrophoresis power is on
- `step3.2_state_m2`: The state where bromocresol green is migrating into gel
- `step3.2_state_m3`: Bromocresol green has migrated into gel approximately 0.5-1 cm, the power of the electrophoresis has been turned off.
- `step3.2_state_m4`: A glass plate has been placed on top of the gel, bromocresol green has migrated less than 2/3 of the gel length.

The process given above is a loop since `step3.2_state_m4` precedes `step3.2_state_m1`. `step3.2_state_m4` differs with `step3.2_state_final` in that the bromocresol green has migrated to the targeted amount in the latter state. `step3.2_state_m3` precedes both `step3.2_state_m4` and `step3.2_state_final`. An instance of **Measure** could be used to track the amount that bromocresol green has migrated.

6.4.1 Ontology Queries using SPARQL

We have used SPARQL to extract some domain knowledge about the experimental procedure of Alkaline Agarose Gel Electrophoresis from our framework. Figures 6.6, 6.7, 6.8, and 6.9, show the true power of knowledge representation by automatically extracting the essential information that a biochemist would use to perform experimental procedures in a lab. These figures show in a few examples how much information can be mined from such a framework with only one experimental procedure. If all standard experimental procedures in biochemistry [22, 133], for example, are modeled and built, one simply cannot imagine how much time and effort will be saved, knowing all essential information is just a few clicks away. Figure 6.6 returns all devices involved in a state of all steps (1.1, 1.2, 3) and Figure 6.8 shows all of the instruments involved in any state for all steps of the Alkaline Agarose Gel Electrophoresis procedure whereas Figure 6.7 shows a query that returned all materials involved in the procedure. Figure 6.9 shows a query that returned the states of step3 and its substeps, which are concerned with measuring the gel length and returned their target values. The ontology was verified to be consistent using the HermiT 1.3.8.3 reasoner [141].

SPARQL Queries

Query1. Return all devices involved in a state of all steps (1.1, 1.2, 3)

```
SELECT ?step ?state ?item
WHERE { ?step rdf:type :Step.
?step :hasState ?state.
?state :involves ?item.
?item rdf:type :Device}
```

Query2. Return all materials involved in all steps

```
SELECT ?step ?state ?item
WHERE { ?step rdf:type :Step.
?step :hasState ?state.
?state :involves ?item.
?item rdf:type/rdfs:subClassOf :Material}
```

Query3. Return all instruments involved in all steps

```
SELECT ?step ?state ?item
WHERE { ?step rdf:type :Step.
?step :hasState ?state.
?state :involves ?item.
?item rdf:type/rdfs:subClassOf :Instrument}
```

Query4. Which states of step 3 and its substeps measure the gel length, and what is the target value?

```
SELECT ?step ?state ?x
WHERE {
:step3 ^:subStep ?step.
?step :hasState ?state.
?state :involves :gel.
:gel :hasMeasure/:hasNumValue ?x}
```

6.5 Remarks

In this chapter, we have proposed a framework that describes the relations and steps of experimental procedures. This framework will enrich the knowledge based systems with

necessary information about experimental procedures that a scientist would automatically access such as instruments (e.g., laboratory centrifuge) and materials (e.g., buffers). Most importantly, this approach is an important step toward our ultimate goal to analyze biomedical articles. This work will be publicly available for the research community to enhance and expand upon. Such a work could be beneficial for various genres that have similar procedure-oriented characteristics. For future work, we also aim to expand our work by incorporating existing ontologies that are essential to this domain such as the ontology for units of measure [127] and the materials ontology [8]. Certain theoretical ontological modelling of states and empirical observations in science can be fruitfully incorporated into our ontology in the future [104].

step	state	item
step1.2	step1.2_state1	device1
step1.2	step1.2_state2	device1
step1.2.1.1	step1.2.1.1_state1	boiling-waterBath
step1.2.1.1	step1.2.1.1_state2	boiling-waterBath
step1.2.2.1	step1.2.2.1_state1	microwaveOven
step1.2.2.1	step1.2.2.1_state2	microwaveOven
step3.2	step3.2_state1	electrophoresis
step3.2	step3.2_state_m3	electrophoresis
step3.2	step3.2_state_m1	electrophoresis
step3.2	step3.2_state2	electrophoresis
step3.2	step3.2_state_m4	electrophoresis
step3.2	step3.2_state_m4	glass_plate

Figure 6.6: Result of Query1: Extract all devices involved in all steps of the Alkaline Agarose Gel Electrophoresis procedure

step	state	item
step1.2.1.1.1	step1.2.1.1.1_state2	h2o1
step1.1.1.1	step1.1.1.1_state1	kimwipe1
step1.1	step1.1_state1	h2o1
step1.2.2.1.1	step1.2.2.1.1_state2	h2o1
step1.2.2.1.2	step1.2.2.1.2_state1	item1
step3.2	step3.2_state_m2	bromocresol_green
step3.2	step3.2_state2	bromocresol_green
step1.2.1.1.1	step1.2.1.1.1_state1	item1
step3.2	step3.2_state_m1	bromocresol_green
step1.2.1.1.1	step1.2.1.1.1_state2	item1
step1.2.1.1.2	step1.2.1.1.2_state1	item1
step1.2.1.1	step1.2.1.1_state1	item1
step1.2.1.1	step1.2.1.1_state2	item1
step1.2	step1.2_state1	item1
step1.2	step1.2_state2	item1
step1.2.2.1	step1.2.2.1_state1	item1
step1.2.2.1	step1.2.2.1_state2	item1
step1.1	step1.1_state1	agarose1
step1.2.2.1.1	step1.2.2.1.1_state1	item1
step1.1	step1.1_state2	step1.1_mixture
step1.2.2.1.1	step1.2.2.1.1_state2	item1

Figure 6.7: Result of Query2: Return all materials involved in all steps of the Alkaline Agarose Gel Electrophoresis procedure

step	state	item
step1.2.2.1.2	step1.2.2.1.2_state1	container1
step1.2.1.1.1	step1.2.1.1.1_state1	container1
step1.2.1.1.1	step1.2.1.1.1_state2	container1
step3.1	step3.1_state1	container3
step3.1	step3.1_state2	container3
step1.2.1.1.2	step1.2.1.1.2_state1	container1
step1.2.1.1	step1.2.1.1_state1	container1
step1.2.1.1	step1.2.1.1_state2	container1
step1.2	step1.2_state1	container1
step1.2	step1.2_state2	container1
step1.2.2.1	step1.2.2.1_state1	container1
step1.2.2.1	step1.2.2.1_state2	container1
step1.1	step1.1_state1	container1
step1.2.2.1.1	step1.2.2.1.1_state1	container1
step1.1	step1.1_state2	container1
step1.2.2.1.1	step1.2.2.1.1_state2	container1
step3.2	step3.2_state_m3	electrophoresis
step3.2	step3.2_state_m4	glass_plate
step3.2	step3.2_state_m4	electrophoresis
step3.2	step3.2_state2	electrophoresis
step3.2	step3.2_state_m1	electrophoresis
step1.2.1.1	step1.2.1.1_state1	boiling-waterBath
step1.2.1.1	step1.2.1.1_state2	boiling-waterBath
step1.2	step1.2_state1	device1
step1.2	step1.2_state2	device1
step1.2.2.1	step1.2.2.1_state1	microwaveOven
step1.2.2.1	step1.2.2.1_state2	microwaveOven
step3.2	step3.2_state1	electrophoresis

Figure 6.8: Result of Query3: Extract all instruments involved in all steps of the Alkaline Agarose Gel Electrophoresis procedure

step	state	x
step3.2	step3.2_state_m4	"2/3"^^<http://www.w3.org/2000/01/rdf-schema#Literal>
step3.2	step3.2_state_m2	"2/3"^^<http://www.w3.org/2000/01/rdf-schema#Literal>
step3.2	step3.2_state2	"2/3"^^<http://www.w3.org/2000/01/rdf-schema#Literal>

Figure 6.9: Result of Query4: Return which states of step3 and its substeps that measure the gel length and what is the target value

Chapter 7

Rhetorical Moves Revisited and the System as a Whole

This thesis focuses on the real world application of scientific writing and on determining rhetorical moves, an important step in establishing the argument structure of biomedical articles. Using the observation that the structure of scholarly writing in laboratory-based experimental sciences closely follows laboratory procedures, we examine most closely the Methods section of the texts and adopt an approach of identifying rhetorical moves that are procedure-oriented. We have proposed earlier in Chapter 4 a VerbNet-like frame semantics with an effective set of semantic roles in order to support the analysis. These components are designed to support a computational model of appropriate rhetorical moves for this domain. Our work also contributes to the understanding of argument-related annotation schemes which is described in Chapter 5. In particular, we conduct a detailed study with human annotators to confirm that our selection of semantic roles is effective in determining the underlying rhetorical structure of existing biomedical articles in an extensive dataset. The annotated dataset that we produce provides the important knowledge needed for our ultimate goal of analyzing biochemistry articles. In this chapter, we revisit the Rhetorical Moves proposed for the model and present as well a picture of the overall processing of a biochemistry article, along with a discussion of some possible uses for that processing. Section 7.1 describes rhetorical moves in biochemistry articles. Section 7.2 describes our overall framework structure and Section 7.3 describes further applications.

7.1 Rhetorical Moves in Biochemistry Articles

We begin by summarizing what we have proposed so far for the Rhetorical Moves component of our model. Various studies have used recurrent patterns of text organization called *rhetorical moves* (i.e., text segments that are rhetorical and perform specific communicative goals) to analyze argumentative organization of texts manually [149] or automatically [154]. Swales' CARS model targets the Introduction section[1] of scientific articles. Teufel's interests are concentrated on rhetorical moves associated with defining the research space and suggesting the knowledge claims for computational linguistics and chemistry articles [151]. Kanoksilapatham [79] adds to these works by providing the first comprehensive set of rhetorical moves for complete biochemistry articles.

Our goal is to provide a computational model for Kanoksilapatham's descriptive rhetorical move taxonomy. Our research agenda is to design algorithms which would produce a representation of rhetorical moves in a biochemistry article. Initially, our focus is on the Methods section of the taxonomy since this provides a description of the procedures followed in the experiment and the analysis of the results of the experiment thereby giving a framework for analyzing the moves in the remainder of the article. Because the experimental process is procedural, the moves tend to follow the verbs describing the steps in the experimental process. In other words, argumentation structure and scientific method both consist of rhetorical moves and experimental process, respectively. When a scientist describes her/his method in the written article, it contains a list of experimental steps which are described by verbs (actions). These verbs evoke (initiate) the rhetorical moves in the writing. To understand the moves, we need information about the semantic roles associated with these procedural verbs which is described in Chapter 4.

Table 7.1 shows our developed rhetorical moves used to describe experimental procedures in the biochemistry articles. These moves have the most frequent occurrence among moves based on our observation and analysis of 105 articles in biochemistry articles, as described in Chapter 3. Since argumentation is constructed through moves in the text, these moves play crucial roles for argumentation analysis in the text. We found that there is a parallel between the steps of experimental procedures and the rhetorical moves in the writing. In other words, the experimental procedures are mirrored by the rhetorical moves in text. In Chapter 4, we described the Semantic Roles and discussed as well the implementation of this component of the model, the natural language processing methods used and some validation of this part of the analysis (using the annotated corpus as the gold standard). We discuss in Section 8.3 some possible steps forward with an independent

[1]Experimental articles in the biomedical sciences are normally organized in the IMRaD style: Introduction, Methods, Results, and Discussion.

Move type	Definition
Description-of-method	Concerned with sentences that describe experimental events.
Appeal-to-authority	Concerned with sentences that discuss the use of well-established methods.
Background information	Concerned with all background information for the experimental events such as "method justification, comment, or observation, exclusion of data, approval of use of human tissue" as defined by Kanoksilapatham (2003).
Source-of-materials	Concerned with the use of certain biological materials in the experimental events.

Table 7.1: Rhetorical moves in the Methods sections of biochemistry articles

validation of the Rhetorical Moves component of the analysis. In the section that follows, we begin to sketch how an overall system which identifies the argument structure of a biochemistry article may be assembled, together with a sense of some of the uses to which this overall system may be put.

7.2 The Overall Structure of Our Framework

Our proposed framework consists of several components: Frame semantics, semantic role labelling, ontology for experimental procedures, and rhetorical move labelling.

So far in this thesis we have described several of the key components that make up our proposed computational model for analyzing biochemistry texts. In this section, we discuss how these components might come together into one algorithm for producing a representation of the argument structure of one these documents.

We begin by reflecting on the possible uses of the automated analysis of biochemistry texts. Some examples include: a) producing a summary of the text b) supporting question/answering about the content of the text. Before these uses can be applied, text processing would occur. The input would be the Methods section of a particular biochemistry text and the output would be a representation indicating the underlying rhetorical moves and verb-based frames with accompanying semantic roles and fillers. Ontologies may be consulted in order to build the representation (for example to enable additional fillers for one of the verb-based semantic frames). Figure 7.1 shows the overall structure

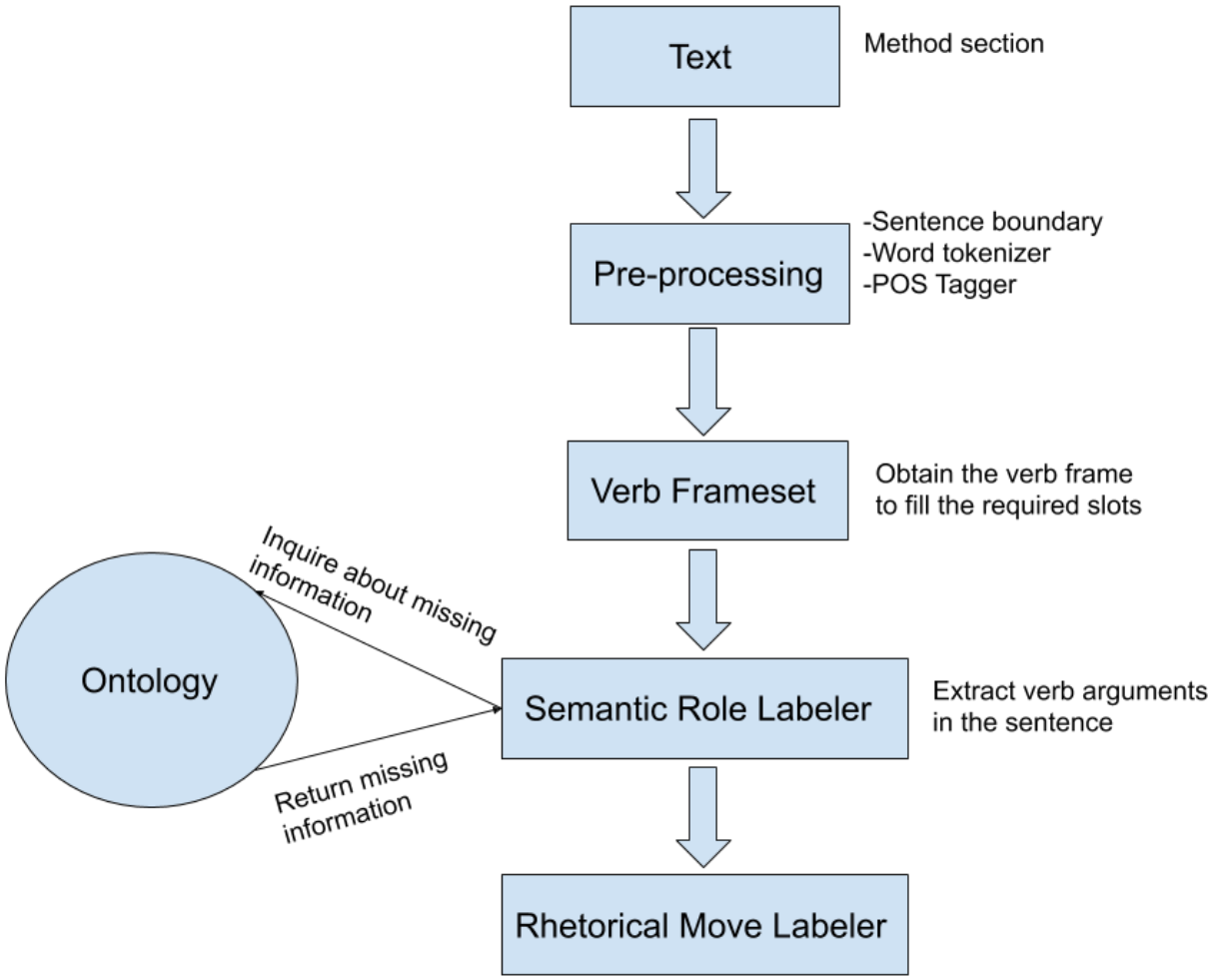

Figure 7.1: The pipeline for our overall framework

of our framework and Algorithm 1 shows the steps of each developed component of our framework. Since it is the knowledge from the frames and verbs that helps to characterize the rhetorical move, we propose performing that stage of the analysis first. Various methods can also be introduced to assist in the identification of these moves, including locating certain cue phrases as signalling a certain move. It is also possible to use machine learning in order to progressively learn which elements arise typically in these texts. The idea is that certain features would contribute most to a certain class of rhetorical move and this can be identified. One might in fact imagine a rule-based script using cue phrases to perform some of the processing.

```
Data: Experimental procedure sentences
Result: List of sentences marked up with information about verbs, semantic roles
        and rhetorical moves
initialization;
while not at end of this document do
    process sentence;
    if verb identified then
        go to next step;
        identify the frameset;
        find all associated verb-arguments;
        if all roles identified then
            go to next step;
            identify rhetorical move for that particular sentence using the knowledge
             from previous steps ;
        else
            use ontologies for experimental procedures to retrieve implied roles;
        end
    else
        go back to the beginning of while loop;
    end
    go back to the beginning of while loop;
end
```

Algorithm 1: Inputs, outputs and steps of the overall framework

7.2.1 Preliminary Validation of the Rhetorical Moves with a Brief Description of the Process and Results

It is our intention to implement the entire process and then to validate it as effective, relative to the annotations provided by our human expert. We discuss how this might proceed in greater detail in the Future Work section of our last chapter.

In this subsection, we describe an preliminary effort to implement the analysis of rhetorical moves which could complement the analysis of semantic roles (4) to get us closer to fully automated system.

The input for this process is a list of sentences from the Methods section. There are four main steps. The first is to identify the verb, through the tagging of the sentence using the GENIA tagger [165] (mentioned in Chapter 3); the second step is to send the verb to a Frame Initiator (a method designed to retrieve the frame associated with the verb from a repository of all frames for all verbs (Appendix F)); the entire structure of the frame is obtained and checked against the input sentence. If this is successful, the third step is to perform semantic role labelling using the model described in Chapter 4, to tag each of the tokens in the sentence with these roles. Once these are retained, we check to make sure that all the roles required for the frame for that verb have been identified. If there are any roles missing, an optional step is to then consult the ontology; for this we use a SPARQL wrapper (querying the ontology with pre-set commands (e.g. checking for Instrument of verb)). At this point the sentence is passed to the rhetorical move labeller, as the fourth step of the process.

How this rhetorical move labeller works is as follows. There are different features to consider. The first is the verb. For example, if the verb is "purchased', it could be potentially counted as Source of Material. If certain semantic roles like location correlated with that verb appear, then this would increase the score for that interpretation of the rhetorical move. This is the second feature considered. As another example, if the sentence discusses Protocol Detail Information such as time, buffer, then this is called Description of Method. The third feature considered are cue phrases, such as "according to", which may also help to identify the rhetorical move (e.g. as Established Method).

At this point, all the sentences of the text would be labelled with rhetorical moves and we need to validate whether we have performed this analysis effectively. The validation is against the gold standard annotation by experts discussed in Chapter 5. The ideal is to use the same dataset used for the validation of semantic roles, as in Chapter 4. For this initial validation of rhetorical moves, we use instead a small set of sentences for which we know the ground truth labelling of rhetorical moves and their associated verb frames. If our analysis agrees with that of the annotators, then we are performing well. To date, we

have inspected by hand that our precision and recall are good. For future work, we need to make use of a larger set and to properly record all the metrics which confirm that our algorithms are performing well.

Examples of the Input and the Output from our Framework

To show a complete Method section being analyzed by our framework would be desirable, but this would require many more verb frames than this thesis work has generated. Instead, to illustrate the computation of rhetorical moves done by the framework, we show four sentences from our dataset as input to the pipeline shown in Figure 7.1. Each sentence corresponds to a specific rhetorical move. The results of the main steps of our proposed framework pipeline are provided. This research work will continue: first, a larger set of verb frames will be developed, and second, some of the algorithms in the pipeline may require more distinct features for each rhetorical move.

Example 1: Description of the method

"Array-generated oligos were amplified four cycles."

Pre-processing stage:
POS tagging: Array-generated/JJ oligos/NNS were/VBD amplified/VBN four/CD cycles/NNS ./.
Chunking: [NP Array-generated oligos] [VP were amplified] [NP four cycles].

Frame Initiator:
From the above POS tagging process, the verb "amplified/VBN" is determined, so this sentence is checked against the set of frames for the verb "amplify" and we have the following match: **Verb Frame:** NP VP NP

Semantic Role Labelling:
Patient: Array-generated oligos
Predicate: were amplified
Protocol detail (Repetition): four cycles.

Rhetorical Move Labeller:
First, the verb "amplifiy" is one of the verbs in the list of verbs that are associated with experimental procedures as defined in Appendix E. Our algorithm would add weight to this being a candidate as a sentence that discusses an experimental procedure. The second step is to determine whether there are cue phrases that evoke certain rhetorical moves such as "according to" or "as described by". Our algorithm didn't detect either in this sentence. The third step is to identify the associated semantic roles from the previous

stage **Semantic Role Labelling** which gave the following pattern: "**Patient Predicate Protocol detail (Repetition)**". So, as it can be seen from the above pattern, the **Protocol detail (Repetition)** information exists in the sentence, so our algorithm adds further weight to this as discussing an experimental procedure and finally labels it as: "**Description of the method**"

Example 2: Appeal to authority

"sCD39 transfected stable HighFive insect cells were cultured as described by Chen and Guidotti [25]."

Pre-processing stage:
POS tagging: sCD39/JJ transfected/JJ stable/JJ HighFive/NNP insect/NN cells/NNS were/VBD cultured/VBN as/IN described/VBN by/IN Chen/NNP and/CC Guidotti/NNP -LSB-/VBZ 25/CD -RSB-/NNS ./.
Chunking: [NP sCD39] [VP transfected] [NP stable HighFive insect cells] [VP were cultured] [SBAR as] [VP described] [PP by] [NP Chen and Guidotti] [[NP 25]] .

Frame Initiator:
From the above POS tagging process, the verb "cultured/VBN" is determined, so this sentence is checked against the set of frames for the verb "cultured" and we have the following match: **Verb Frame:** NP VP SBAR PP

Semantic Role Labelling:
Theme: sCD39 transfected stable HighFive insect cells
Predicate: were cultured
Instrument (Reference): as described by Chen and Guidotti [25].

Rhetorical Move Labeller:
First, the verb "culture" is one of the verbs in the list of verbs that are associated with experimental procedures as defined in Appendix E. Our algorithm would add weight to this being a candidate as a sentence that discusses an experimental procedure. The second step is to determine whether there are cue phrases in the above sentence. Our algorithm detects the cue phrase "as described by" followed by a citation to other work "Chen and Guidotti [25]" in this sentence, so the algorithm would add weight to this being a candidate for appeal to authority. The third step is to identify the associated semantic roles from the previous stage **Semantic Role Labelling** which gave the following pattern: "**Theme Predicate Instrument (Reference)**". So, as it can be seen from the above pattern, the **Instrument (Reference)** information exists in the sentence, so our algorithm adds further weight to this as discussing an appeal to authority and finally labels it as: "**Appeal to authority**".

Example 3: Source of materials

"All reagents were purchased from Sigma Aldrich (Oakville, ON, Canada)."

Pre-processing stage:
POS tagging: All/DT reagents/NNS were/VBD purchased/VBN from/IN Sigma-Aldrich/NNP -LRB-/-LRB- Oakville/NNP ,/, ON/NNP ,/, Canada/NNP -RRB-/-RRB- ./.
Chunking: [NP All reagents] [VP were purchased] [PP from] [NP Sigma-Aldrich] ([NP Oakville] , [NP ON] , [NP Canada]) .

Frame Initiator:
From the above POS tagging process, the verb "purchased/VBN" is determined, so this sentence is checked against the set of frames for the verb "purchased"[2] and we have the following match: **Verb Frame:** NP VP PP

Semantic Role Labelling:
Theme: All reagents
Predicate: were purchased
Location: from Sigma Aldrich (Oakville, ON, Canada).

Rhetorical Move Labeller:
First, the verb "purchase" is one of the verbs in the list of verbs that are associated with source of materials. Our algorithm would add weight to this being a candidate for a sentence that describes the source of materials. The second step is to identify the associated semantic roles from the previous stage **Semantic Role Labelling** which gave the following pattern: "**Patient Predicate Location**". So, as it can be seen from the above pattern, the **Location** information appears in the sentence, so our algorithm would add further weight to this as discussing a purchase of materials and finally labels it as: "**Source of materials**"

Example 4: Background information

"We used CHOK1cells as an alternative host in these studies."

Pre-processing stage:
POS tagging: We/PRP used/VBD CHOK1cells/NNS as/IN an/DT alternative/JJ host/NN in/IN these/DT studies/NNS ./.
Chunking: [NP we] [VP used] [NP CHOK1cells] [PP as] [NP an alternative host] [PP in] [NP these studies] .

[2]In this case, the verb "purchased" is not in our list of frames. However, we will use the chunk parser output to serve our purposes here for demonstration, as this work will be developed further as discussed in Chapter 8.

Frame Initiator:
From the above POS tagging process, the verb "used/VBN" is determined, so this sentence is checked against the set of frames for the verb "used"[3] and we have the following match:
Verb Frame: NP VP NP PP PP

Semantic Role Labelling:
Agent: we
Predicate: used
Theme: CHOK1cells.
Patient: as an alternative host.
Protocol detail (Condition): in these studies.

Rhetorical Move Labeller:
First, the verb "used" is one of the verbs in the list of general verbs that could be used to describe background information such as method justifications or comments. Our algorithm would add weight to this being a candidate for a sentence that discusses background information. The second step is to determine whether there are cue phrases in the above sentence related to background information that shows some discussions or observations such as "as an alternative" or "unfortunately". This sentence contains "as an alternative host". The third step is to identify the associated semantic roles from the previous stage **Semantic Role Labelling** which gave the following pattern: "**Agent Predicate Theme Patient Protocol detail (Condition)**" So, as it can be seen from the above pattern, the semantic role **Agent** is present in the sentence. This indicates that a discussion or some reasoning about a particular choice that the authors made in their experiments took place. So, our algorithm would add further weight to this as background information and finally labels it as: "**Background information**"

7.3 Further Applications

While our primary long-term goal is to automatically mine the argumentation in a biochemistry article, we now return to clarify how the output of our argument analysis of a text can be used for tasks such as summarizing or question/answering. The first point is that different parts of the representation may be required, depending on the task. To summarize at a high level, the list of rhetorical moves may be most useful, though to actually generate the language of the summary, one would need to present the underlying verb of

[3]In this case, the verb "used" is not in our list of frames. However, we will use the chunk parser output to serve our purposes here for demonstration, as this work will be developed further as discussed in Chapter 8.

Beads with bound proteins were washed six times (for 10 min under rotation at 4C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography.

Figure 7.2: Input for summarization based on our framework

Beads with bound proteins were washed and proteins harvested, separated, and analyzed.

Figure 7.3: Ouput for summarization based on our framework

the associated verb frame and also generate text which covers some of the semantic role fillers. For question/answering, it may be especially useful to consult the ontology in order to introduce knowledge associated with each rhetorical move.

An example of a summarization task is given above. Figures 7.2 and 7.3 show the input and output of a sample text. The processes of our framework involve first identification of all verbs involved in the sentence. Since the main idea of summarization is to condense the information presented in the sentence to its essential. So, the step of determining the frameset for each verb is incorporated since it provides a list of mandatory (e.g., theme, and agent) and optional (e.g., time, and buffer) semantic roles for a specific verb. Then the step for finding all semantic roles that exist in the sentence is employed. Once all semantic roles are identified for each verb, the system can simply omit non-essential information by leaving out optional roles such as time, buffer, condition and temp.

Chapter 8

Discussion, Conclusion and Future Work

In this chapter, we first of all step back to reflect on the lessons learned in this thesis and the primary contributions for researchers working on related problems. We then draw out some of the main benefits derived from our current solution, in comparison with other models in the literature. After this Discussion section, we move on to reveal the primary conclusions of this thesis, providing as well a detailed list of each of the primary contributions that we offer. The final section of this chapter is devoted to a reflection on many of the steps forward that are possible with this research, suggesting possible starting points and explaining the potential value of these additional threads of research.

8.1 Discussion

The first topic of this section is advice to other researchers who are trying to design natural language processing algorithms for scientific texts that focus on representing the underlying arguments. We elaborate in particular on the lessons we learned when attempting to produce annotated texts to use as a gold standard for validating our approach: the challenges of working with human annotators and the most productive steps forward. The second topic is our discussion of the value of our proposed model, in comparison with those of other researchers. Here we choose to describe key advances provided by our solution, in comparison with that of Kanoksilapatham [79] (which has been identified already as comprehensive but largely descriptive) and to contrast with some of the most advanced efforts to construct argument structure in biomedical texts, the work of Nancy Green [62, 63].

8.1.1 Challenges Observed and Lessons Learned for Text Analysis of Scientific Text

Although, in this thesis, our focus is on one discipline, biochemistry articles, we believe that the work presented in this dissertation could potentially benefit researchers in various disciplines such as chemistry, ecology, pharmaceutical science, environmental science, and biology. In spite of our focus on the Method section, other sections in biochemistry articles could also benefit from our developed annotation procedures based on the concept of verb-centric frame semantics. The annotation procedures and the creation of guidelines were the most essential components of our study. These steps include: the preparation of the dataset, involvement of domain experts, recruitment of annotators, training the annotators on using the annotation tools, and a commitment to answer their questions and meeting with them regularly to ensure they are on track. The annotation study involved more than 15 students; however, only a few of them continued with us up to the writing of this thesis. The annotation process also has a financial cost. However, fruitful results came after the hard work. The first email sent to recruit annotators in this study was November 2017 and until the writing of this thesis, annotators were working on labelling biochemistry articles. I have summarized lessons learned in the following points, of possible use for other researchers:

- It is valuable to take the first step yourself, working with a corpus of articles in your domain, to get a sense of the appropriate choice for rhetorical moves and for semantic roles and frames.
- The next step would be to have domain experts help to validate whether your understanding of the articles was acceptable or not. Some kind of iteration between you and the experts may lead to refinement of the proposed knowledge representation.
- That second step may well lead you to conclude that some additional Knowledge Representation may be needed, to fill in the gaps of what is said vs. what is to be interpreted (e.g. our ontologies).
- Pay a lot of attention to developing appropriate guidelines for the annotators and allow them to discuss between themselves in order to help with agreement.
- As annotators are being trained, put very good example articles in front of them and try to give all the annotators the same article to begin with, so that the group can collectively discuss in order for everyone to see how the texts should best be labelled.

- After this, come up with an executive decision of what the labelling should be, to inform the annotators about the best consensus, as part of their training.
- Train the annotators to consider carefully the appropriate span for the annotation as well.
- In order to have better agreement without bias, ask annotators not to discuss any further and then to simply field generic questions through Slack, with other queries directed to the researcher at this point.

8.1.2 Comparisons with Key Related Work

Our work is a step forward from the work proposed by Kanoksilapatham [79]. Essentially, Kanoksilapatham advanced Swales' approach to move analysis by developing a framework that combines his original CARS model with the use of Biber's multidimensional analysis [14] to enrich the model with additional information about linguistic characteristics. However, Kanoksilapatham's extension to Swales' move analysis study is merely a descriptive analysis about rhetorical moves without defining an explicit method for analyzing and recognizing these moves in texts. In this thesis, we have advanced Kanoksilapatham's move analysis by providing a knowledge representation framework based on the *verb-centric frame semantics*. Kanoksilapatham's [80] dataset and the number of annotators, one PhD student, involved was relatively small compared to our developed dataset and number of annotators.

Our work on annotation, described in Chapter 5, is also similar to the one proposed by Green [62]. Her goal was to develop an annotated corpus of biomedical genetics research articles. Green [63] specified a set of argumentation schemes to label scientific claims. Green's annotation involved the identification of premises, conclusions of an argument as well as its type of scheme. Based on the analyses of various genetics research articles, Green specified 10 argumentation schemes that are semantically different. Some of these schemes were new and had not previously been proposed. Furthermore, the specification of argumentation schemes was used to create annotation guidelines. Then, these guidelines were evaluated in a pilot study based on participants' ability to recognize these schemes by reading the guidelines. However, based on the pilot study, the results showed a variation in performance since there were two groups of participants (i.e., undergraduate students and researchers). In contrast, our annotation was primarily focused on the rhetorical moves and the identification of core aspects of sentences (i.e., verbs and their arguments). We have developed an annotation scheme which consists of a set of semantic roles. Some of them are well-established in the literature and others are suggested by domain experts

and required in this domain. We also have used a set of well-established rhetorical moves that are suggested by well-respected researchers in the literature [149, 80]. In our case, our annotation guidelines were developed in an in-depth fashion. They were used with domain experts to create an annotated corpus of experimental procedures and we have achieved substantial agreement in identifying semantic roles. Our overall representation of the arguments includes this deeper level of detail not covered by Green.

8.2 Conclusion

The main focus of this thesis is rhetorical moves in biochemistry articles. Kanoksilapatham [79] has provided a descriptive theory of rhetorical moves that extends Swales' [149] CARS model to the complete biochemistry article. We have developed and described a computational model of Kanoksilapatham's descriptive theory in Chapter 7. Our hypothesis is that recognizing and detecting rhetorical moves would provide important information to our argumentation analysis framework, and that the Method sections in biochemistry articles contain moves which can be correlated with the author's experimental procedures. These moves can be used to determine salient information about the elements of the article's argumentative structure (e.g., premises) and can contribute to the overall understanding of the author's scientific claims. A key aspect of our hypothesis is that development of a frame-based knowledge representation can be based on the semantics of the verbs associated with these procedures. This representation can provide detailed knowledge for understanding these rhetorical moves, which will in turn facilitate analysis of argumentation structure. In other words, we propose that a *procedurally rhetorical verb-centric frame semantics* can be used to obtain a sufficiently deep analysis of sentence meaning which was described in Chapter 4. We have proposed an extension to the general semantic roles that are suited for experimental procedures which is also described in Chapter 4. We also have developed a corpus of Method sections that have been marked up for rhetorical moves and semantic roles, described in Chapter 5. We also have developed a prototype ontology that provides experimental procedure knowledge for the biochemistry domain since the writing style of this genre tends to occasionally omit important information, described in Chapter 6. Our computational model employed machine learning to build its models for the semantic roles, validated against a gold standard reflecting the annotation of these texts by human experts, which is described as well in Chapter 5. We provided significant insights into how to derive these annotations, and as such have contributions as well to the general challenge of producing markups in the domain of biomedical science documents, where specialized knowledge is required.

8.2.1 Central Contributions of the Thesis

Our central contributions of the thesis are:

1. The creation of an annotated dataset marked up with both information about verbs and their arguments and rhetorical moves, described in Chapter 5
2. The development of the semantic roles scheme, described in Chapter 4
3. The creation of frames for procedural verbs and their definitions and usages, described in Chapter 4
4. The Annotation procedures and guidelines, described in Chapter 5
5. The development of prototype ontology based knowledge representation, described in Chapter 6

The benefits of these central contributions can be understood in greater detail, as follows. Annotation guidelines will assist other researchers in deciding how to create their datasets and to achieve annotations. The annotated dataset is very helpful for the domain in general useful for summarizing, Q/A etc. and extracting semantic or syntactic information or to train shallow parsing. This is also support for community of argumentation researchers, as a new labelled dataset and for future validation efforts. As for the development of semantic roles, our approach was based on the knowledge representation of frames; others in different domains could benefit from the use of independent semantic roles or use the ones we propose for any domains that have procedure-oriented aspects. The ontology development was first made very specific, for our task and then we realized its possible use in other domains; it is now sufficiently generic to be employed in other applications as well. In general, we also demonstrate the value of introducing ontologies into the NLP solutions for scientific articles, which other researchers may want to consider when developing their analyses of these kinds of documents. The SPARQL queries supported by the ontology can be helpful for other researchers as well. The use of instrument we discovered to be very frequent for scientific articles and we now have a deeper knowledge representation of this concept worked out in some detail. We discovered the need to go beyond VerbNet and FrameNet for our semantic representations, to be able to process in a domain-specific way; other researchers working on analyzing scientific writing should now be aware of the importance of incorporating a step like this.

8.3 Future Work

There are several possible steps forward with this work, some of which we explored briefly ourselves while assembling our overall model.

8.3.1 Verb Frames

We have developed a set of frames for frequent procedural verbs (e.g., "digest") in our analyzed data set. Our aim was to extend the VerbNet project [137] by providing syntactic and semantic frames for the procedural verbs described in Chapter 4. As future work, we could develop a larger set of frames for the procedural verbs in biochemistry as an extension to lexical resources such as the VerbNet project. We need also to examine how frequent these frames in larger corpora which require manual validation and analysis. Following the VerbNet project [137], we plan to classify verbs that share similar activity or common meaning "sense" in a particular class such as "harvest" and "collect". We also could incorporate some verbs (e.g., "denatured") in some VerbNet classes such as "class 10" which is related to "verbs of removing" or verbs like "carboxymethylated" in "class 20" which is referred to as "verbs of contact".

8.3.2 Analysis of Other Sections in Biochemistry Articles

Our long-term plan, beyond the scope of this thesis, is to analyze all sections of biochemistry articles (IMRaD) and build a framework that is capable of identifying key aspects of text (e.g., argumentative elements). However, this is a huge task and it involves many sub-projects that could be ideas for several PhD theses. Since we have focused on the Method section in this thesis, the analysis of other sections (e.g., discussion and results) in biochemistry articles is a potential future work that would align with the work in this thesis and would provide an important contribution for analyzing the overall structure of argumentation in biochemistry articles.

8.3.3 Re-Training Our SRL System on a Large Annotated Dataset

We noted that the accuracy of our SRL system suffered due to two important aspects: The training data is small, and the annotated dataset contains blemishes. Some verb arguments are either mislabeled (e.g., labeled as "Theme" where it should be an "Instrument") or left

without any label. Our dataset as we described in Chapter 4 is small compared to datasets such as CONLL2005, which includes over 39K sentences, while our dataset contains only 8778 sentences. Future work will entail further annotation efforts which includes annotating more data and cleaning the resulting annotations.

8.3.4 Rhetorical Move Labelling

As future work, we aim to develop a neural network model to classify rhetorical moves into their proper category. The model will be trained on using the marked up dataset with information associated with each verb (semantic roles) along side with move categories in Chapter 5. Ultimately, we will test the model on an unseen test set to calculate the accuracy of the model. Due to time and available funds, we could not finish this step.

8.3.5 Automatic System for the Overall Framework Structure

One possible extension of this work is to implement the overall framework structure described in Chapter 7 by writing an algorithm to interconnect components with each other. We have begun to sketch the processing required.

8.3.6 Exploring Other Rhetorical Moves

In working with our annotators, we discovered that it was difficult at times to reach agreement on labelling of the rhetorical moves. In particular, the choice between Description of Method and Appeal to Authority differed between annotators. We note that this aligns with comments made by Kanoksilapatham [80] which stated that these two moves are closely related and could be classified under one main category which is "Describing experimental procdures".

It was valuable, all the same, for us to observe which moves appeared most frequently and this is an item that may merit further attention in the future. We could, for instance, experiment with a slightly different set of rhetorical move choices, with our annotators, to see if the agreement improves and then use this as the basis for our implementation and validation. Or we may be able to examine validations performed with competing sets of rhetorical moves, in order to learn which is most effective. Ultimately, applying the algorithms developed towards the end uses which we discuss in Chapter 7 such as summarization or question answering will shed the most light on what the best solution is here.

8.3.7 Semantic Role Labelling

Although there are many existing neural-net-based semantic role labelling systems available in the literature, these systems have been designed to be trained on the CONLL datasets. One design characteristic of these neural net systems is that the size of the input (in this case, the maximum length of the sentences) must be predetermined. The limit on the number of tokens in a sentence in the semantic role labeler used in this thesis is 100 words [67]. One aspect of sentences contained in the Method section of biochemistry articles is that they can be very long (i.e., more than 100 words). The 100 word limit causes salient sentences that contain potentially important information to be thrown away. So sentences like the following are not available for training and testing or further analysis: "For Y2H mapping experiments FF domain constructs from Prp40 as baits and from Snu71 as prey were produced using the following primers (forward primers shown as codons): ForwardFF1: A TTC CAG CTG ACC ACC ATG AGA AGG ACT AAA GAA GAA, ReverseFF1: GA TCC CCG GGA ATT GCC ATG TGT TTC ATT GTG TTC CT, ForwardFF2: A TTC CAG CTG ACC ACC ATG AAG GAA CAC AAT GAA ACA, ReverseFF2: GAT CCC CGG GAA TTG CCA TGG ATT CTT TCT GAG TGT CG, ForwardFF3: A TTC CAG CTG ACC ACC ATG AAT TAT ACC AGA GAC CGT, ReverseFF3: ATC CCC GGG AAT TGC CAT GAC GTC TGT TGG GCT ATT G, ForwardFF4: A TTC CAG CTG ACC ACC ATG CAA AAT GAG CGT AGG ATA, ReverseFF4: GAT CCC CGG GAA TTG CCA TGC GCT TTC GGC AGT CGG ForwardSnu71II: AA TTC CAG CTG ACC ACC ATG TCC GAG AGA AGC GCG GCA GAG, ReverseSnu71II: GAT CCC CGG GAA TTG CCA TGC TCT GCC GCG CTT CTC TCG GA, ForwardSnu71I: AA TTC CAG CTG ACC ACC ATG GCC AAA GGG AGC GCC AAT ACA, ReverseSnu71I: GAT CCC CGG GAA TTG CCA TGT GTA TTG GCG CTC CCT TTG GC." [47]

This sentence specifies what primers are used and also describes specific sections of the yeast DNA. A biochemist would find the above sentence important especially if she/he would like to reproduce the experiment. Thus, there is a need for a semantic role labelling system that is tailored to the experimental procedures domain. Exploring how to accomplish this is left for future work.

8.3.8 Expanding the Procedure-oriented Ontology

As a future work, we aim to expand our prototype ontology to include various experimental procedures from the manual of standard biochemistry procedures [22]. This expansion is an important step toward our ultimate goal to analyze biomedical articles. We also aim

to expand our work by incorporating existing ontologies that are essential to this domain such as the ontology for units of measure [127] and the materials ontology [8]. Certain theoretical ontological modelling of states and empirical observations in science can be fruitfully incorporated into our ontology in the future [104]. With our case study for gel purification in Chapter 6, we have an understanding of the completeness of our ontology for answering any query about the experimental procedures, for this context. For future work, we can examine and evaluate our prototype ontology with respect to other experimental procedures in order to test its completeness.

8.3.9 Exploring Whether Methods Sections are Central to Scientists

In Chapter 3, we clarified our assumption that Methods sections of biochemistry articles provide the central insights for scientists and thus are best to capture, when depicting the argument structure of the text. It would be possible for future work to continue to confirm that this assumption is well-founded, by displaying our proposed argument analysis of biochemistry articles to a number of scientific researchers, using some kind of user study or survey. This would dovetail with our proposed effort in Section 8.3.2, to examine other components of the articles as well.

8.3.10 Extending the Frame Semantics and Engaging Scientific Authors

It would be valuable to expand upon our verb-centric procedurally rhetorical frame semantics to make the representations even richer. For example, synonmy and hyponymy could be specified as well. Another intriguing idea is to display the representations of the text (rhetorical moves and semantic roles) to biochemistry scientists, to see if this might encourage writing that is inherently more accessible to readers.

8.4 Final Remarks

This genre is very rich in terms of information contained in the research articles and resources available (e.g., the de facto bible) [133] that aid in understanding the steps involved in experimental procedures which make it ideal for further investigation and development.

With the sheer volume of the available biomedical articles thanks to PubMed, many Natural Language Processing (NLP)/computational linguistics researchers have been attracted to this biomedical domain to develop systems which manipulate, retrieve, and extract specific information including: protein-protein interactions (PPI) [87], drug-drug interactions (DDI) [140], gene relationships [74], mining biomedical relations and events [97], and protein-residue associations [124]. We hope to see more involvement from NLP researchers to develop tools that utilize the available biomedical repositories. Our thesis has delivered a significant step forward.

References

[1] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, 2012.

[2] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, 2014.

[3] Mohammed Alliheedi and Robert E. Mercer. Semantic roles: Towards rhetorical moves in writing about experimental procedures. In *Proceedings of the 32nd Canadian Conference on Artificial Intelligence*, pages 518–524, 2019.

[4] Mohammed Alliheedi, Robert E. Mercer, and Robin Cohen. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, 2019.

[5] Mohammed Alliheedi, Robert E. Mercer, and Sandor Haas-Neill. Ontological knowledge for rhetorical move analysis. In *The 20th International Conference on Computational Linguistics and Intelligent Text Processing*. To appear in the Journal of Computacion y Sistemas (CyS), 2019.

[6] Mohammed Alliheedi, Yetian Wang, and Robert E. Mercer. Biochemistry procedure-oriented ontology: A case study. In *The 11th International Conference on Knowledge Engineering and Ontology Development*. To appear in Science and Technology Publications (SCITEPRESS), 2019.

[7] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T.

Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[8] Toshihiro Ashino. Materials Ontology: An infrastructure for exchanging materials information and knowledge. *Data Science Journal*, 9:54–61, 2010.

[9] Olga Babko-Malaya. PropBank annotation guidelines. Retrieved from https://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf, 2005.

[10] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, pages 86–90, 1998.

[11] Catherine A. Ball, Kara Dolinski, Selina S. Dwight, Midori A. Harris, Laurie Issel-Tarver, Andrew Kasarskis, Charles R. Scafe, Gavin Sherlock, Gail Binkley, Heng Jin, et al. Integrating functional genomic information into the Saccharomyces genome database. *Nucleic Acids Research*, 28(1):77–80, 2000.

[12] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Marcus C. Chibucos Bill Bug and, Kevin Clancy, Mlanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Norman Morrison Monnie McGee and, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Jr. Christian J. Stoeckert, Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. The ontology for biomedical investigations. *PLoS ONE*, 11(4):e0154556, 2016.

[13] John F. Barrett, Linda A. Lee, and Chi V. Dang. Stimulation of MYC transactivation by the TATA binding protein in promoter-reporter assays. *BMC Biochemistry*, 6:7, 2005.

[14] Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, 1991.

[15] Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. Automatic claim negation: Why, how and when. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 84–93, 2015.

[16] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the Natural Language ToolKit*. O'Reilly Media, Inc., 2009.

[17] Judith A. Blake, Janan T. Eppig, Joel E. Richardson, Muriel T. Davisson, and the Mouse Genome Database Group. The Mouse Genome Database (MGD): Expanding genetic and genomic resources for the laboratory mouse. *Nucleic Acids Research*, 28(1):108–111, 2000.

[18] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[19] Ian Bogost. The rhetoric of video games. In Katie Salen, editor, *The Ecology of Games: Connecting Youth, Games, and Learning*, The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, pages 117–140. The MIT Press, 2008.

[20] Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, 2014.

[21] Filip Boltužić and Jan Šnajder. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, 2015.

[22] R. F. Boyer. *Biochemistry Laboratory: Modern Theory and Techniques*. Prentice Hall, 2012.

[23] Ryan R. Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M. Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Larisa N. Soldatova, Christian J. Stoeckert Jr., Jessica A. Turner, Jie Zheng, and the OBI consortium. Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*, 1 (Suppl 1):S7, 2010.

[24] Neil G. Brown, D. Nick Morrice, Georges Beaud, Grahame Hardie, and David P. Leader. Identification of sites phosphorylated by the vaccinia virus B1R kinase in viral protein H5R. *BMC Biochemistry*, 1:2, 2000.

[25] Camilla Burnett, Panagiota Makridou, Lindsay Hewlett, and Ken Howard. Lipid phosphate phosphatases dimerise, but this interaction is not required for *in vivo* activity. *BMC Biochemistry*, 5:2, 2004.

[26] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, pages 208–212, 2012.

[27] Anne L. Carenbauer, James D. Garrity, Gopal Periyannan, Robert B. Yates, and Michael W. Crowder. Probing substrate binding to Metallo-β-Lactamase L1 from Stenotrophomonas maltophilia by using site-directed mutagenesis. *BMC Biochemistry*, 3:4, 2002.

[28] Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, (CONLL '05), pages 152–164, 2005.

[29] Lucas Carstens and Francesca Toni. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, 2015.

[30] Fa-An Chao, Aleardo Morelli, John C. Haugner III, Lewis Churchfield, Leonardo N. Hagmann, Lei Shi, Larry R. Masterson, Ritimukta Sarangi, Gianluigi Veglia, and Burckhard Seelig. Structure and dynamics of a primordial catalytic fold generated by *in vitro* evolution. *Nature Chemical Biology*, 9(2):81–83, 2013.

[31] Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 132–139, 2000.

[32] Wei Chen and Guido Guidotti. The metal coordination of sCD39 during ATP hydrolysis. *BMC Biochemistry*, 2:9, 2001.

[33] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[34] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213, 1968.

[35] K. Bretonnel Cohen, Philip V. Ogren, Lynne Fox, and Lawrence Hunter. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 38–45, 2005.

[36] Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical Natural Language Processing*. Natural Language Processing 11. John Benjamins Publishing Company, 2014.

[37] Mélanie Courtot, William Bug, Frank Gibson, Allyson L. Lister, James Malone, Daniel Schober, Ryan R. Brinkman, and Alan Ruttenberg. The OWL of biomedical investigations. In *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions*, 2008.

[38] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105, 2010.

[39] Wayne I.L. Davies, T. Katherine Tamai, Lei Zheng, Josephine K. Fu, Jason Rihel, Russell G. Foster, David Whitmore, and Mark W. Hankins. An extended family of novel vertebrate photopigments is widely expressed and displays a diversity of function. *Genome Research*, 25(11):1666–1679, 2015.

[40] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350, 2007.

[41] Sarah K. Deng, Yi Yin, Thomas D. Petes, and Lorraine S. Symington. Mre11-Sae2 and RPA collaborate to prevent palindromic gene amplification. *Molecular Cell*, 60(3):500–508, 2015.

[42] Premkumar T. Devanbu and Diane J. Litman. Taxonomic plan reasoning. *Artificial Intelligence*, 84(1-2):1–35, 1996.

[43] Andrew Dolbey, Michael Ellsworth, and Jan Scheffczyk. BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In *Second International Workshop on Formal Biomedical Knowledge Representation*, (KR-MED 2006), pages 87–94, 2006.

[44] Michel Dumontier, Christopher J.O. Baker, Joachim Baran, Alison Callahan, Leonid Chepelev, José Cruz-Toledo, Nicholas R. Del Rio, Geraint Duck, Laura I. Furlong, Nichealla Keath, Dana Klassen, James P. McCusker, Núria Queralt-Rosinach, Matthias Samwald, Natalia Villanueva-Rosales, Mark D. Wilkinson, and Robert Hoehndorf. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5(1):14, 2014.

[45] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

[46] Ahmed El Zoeiby, François Sanschagrin, Pierre C. Havugimana, Alain Garnier, and Roger C. Levesque. In vitro reconstruction of the biosynthetic pathway of peptidoglycan cytoplasmic precursor in Pseudomonas aeruginosa. *FEMS Microbiology Letters*, 201(2):229–235, 2001.

[47] Claudia Ester and Peter Uetz. The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochemistry*, 9:29, 2008.

[48] Vanessa Wei Feng and Graeme Hirst. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 987–996, 2011.

[49] Charles J Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

[50] Charles J Fillmore. Topics in lexical semantics. *Current Issues in Linguistic Theory*, 76:138, 1977.

[51] C.J. Fillmore. *Santa Cruz Lectures on Deixis, 1971*. Indiana University Linguistics Club, 1975.

[52] Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, 2015.

[53] Astrid Fleischmann, Michael Darsow, Kirill Degtyarenko, Wolfgang Fleischmann, Sinéad Boyce, Kristian B. Axelsen, Amos Bairoch, Dietmar Schomburg, Keith F. Tipton, and Rolf Apweiler. Intenz, the integrated relational enzyme database. *Nucleic Acids Research*, 32(suppl_1):D434–D437, 2004.

[54] FlyBase Consortium. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Research*, 31(1):172–175, 2003.

[55] Gene Ontology Consortium. The Gene Ontology: Enhancements for 2011. *Nucleic Acids Research*, 40(D1):D559–D564, 2011.

[56] Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, 2014.

[57] Olga Gladkova. *Identification of epistemic topoi in a corpus of biomedical research articles*. PhD thesis, University of Waterloo, 2011.

[58] Olga Gladkova, Chrysanne DiMarco, and Randy Allen Harris. What's in a name: Journal titles in the field of epistemic research. *Journal of Argumentation in Context*, 3(3):259–286, 2014.

[59] Olga L. Gladkova, Chrysanne DiMarco, and Randy Allen Harris. Argumentative meanings and their stylistic configurations in clinical research publications. *Argument & Computation*, 6(3):310–346, 2015.

[60] A. Graves, M. Liwicki, S. Fernndez, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

[61] Heather Graves, Roger Graves, Robert Mercer, and Mahzereen Akter. Titles that announce argumentative claims in biomedical research articles. In *Proceedings of the First Workshop on Argumentation Mining*, pages 98–99, 2014.

[62] Nancy Green. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, 2014.

[63] Nancy Green. Identifying argumentation schemes in genetics research articles. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 12–21, 2015.

[64] Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, 2014.

[65] Nancy Green, Rachael Dwight, Kanyamas Navoraphan, and Brian Stadler. Natural language generation of biomedical argumentation for lay audiences. *Argument & Computation*, 2(1):23–50, 2011.

[66] Arthur Hastings. *A reformulation of the modes of reasoning in argumentation*. PhD thesis, Northwestern University, 1963.

[67] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and What's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 473–483, 2017.

[68] Myriam A. Hernández and José M. Gómez. Survey in sentiment, polarity and function analysis of citation. In *Proceedings of the First Workshop on Argumentation Mining*, pages 102–103, 2014.

[69] Annette Højmann Larsen, Aase Frandsen, and Marek Treiman. Upregulation of the SERCA-type Ca^{2+} pump activity in response to endoplasmic reticulum stress in PC12 cells. *BMC Biochemistry*, 2:4, 2001.

[70] Gill Holdsworth, Daniel A. Osborne, TrucChi Thi Pham, James I. Fells, Gillian Hutchinson, Graeme Milligan, and Abby L. Parrill. A single amino acid determines preference between phospholipids and reveals length restriction for activation of the S1P4 receptor. *BMC Biochemistry*, 5:12, 2004.

[71] Hospice Houngbo and Robert Mercer. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, 2014.

[72] Chung-Chi Huang and Zhiyong Lu. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144, 2016.

[73] Lawrence Hunter and K. Bretonnel Cohen. Biomedical language processing: What's beyond PubMed? *Molecular Cell*, 21(5):589–594, 2006.

[74] Junguk Hur, Arzucan Özgür, Zuoshuang Xiang, and Yongqun He. Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *Journal of Biomedical Semantics*, 3:18, 2012.

[75] Syeed Ibn Faiz and Robert Mercer. Extracting higher order relations from biomedical text. In *Proceedings of the First Workshop on Argumentation Mining*, pages 100–101, 2014.

[76] Nancy Ide. Challenges for scientific publication mining. Keynote Speaker presentation at the 32nd Canadian AI Conference, Kingston, Canada, May 2019.

[77] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Research*, 34(suppl_1):D354–D357, 2006.

[78] Juyeon Kang and Patrick Saint-Dizier. Requirement mining in technical documents. In *Proceedings of the First Workshop on Argumentation Mining*, pages 108–109, 2014.

[79] Budsaba Kanoksilapatham. *A corpus-based investigation of scientific research articles: Linking move analysis with multidimensional analysis*. PhD thesis, Georgetown University, 2003.

[80] Budsaba Kanoksilapatham. Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24(3):269–292, 2005.

[81] Manfred Kienpointner. *Alltagslogik: Struktur und Funktion von Argumentationsmustern. Stuttgart*. Stuttgart-Bad Cannstatt : Frommann-Holzboog, 1992.

[82] Johannes Kiesel, Khalid Al Khatib, Matthias Hagen, and Benno Stein. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38, 2015.

[83] Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, (LREC'02), pages 1989–1993, 2002.

[84] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.

[85] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, 2015.

[86] Tomas Klingström, Larissa Soldatova, Robert Stevens, T. Erik Roos, Morris A. Swertz, Kristian M. Müller, Matúš Kalaš, Patrick Lambrix, Michael J. Taussig, Jan-Eric Litton, Ulf Landegren, and Erik Bongcam-Rudlof. Workshop on labortory protocol standards for the Molecular Methods Database. *New Biotechnology*, 30(2):109–113, 2013.

[87] Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(2):S4, 2008.

[88] John D. Lafferty, Andrew McCallum, and Fernando C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, (ICML'01), pages 282–289, 2001.

[89] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[90] M.L. Larson. *Meaning-Based Translation: A Guide to Cross-Language Equivalence*. University Press of America, 1984.

[91] Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, 2018.

[92] John Lawrence and Chris Reed. Combining argument mining techniques. In *Proceedings of the Second Workshop on Argumentation Mining*, pages 127–136, 2015.

[93] John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, 2014.

[94] Douglas B Lenat, Mayank Prakash, and Mary Shepherd. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, 6(4):65–65, 1985.

[95] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, 2012.

[96] Thomas Lippincott, Laura Rimell, Karin Verspoor, and Anna Korhonen. Approaches to verb subcategorization for biomedicine. *Journal of Biomedical Informatics*, 46(2):212–227, 2013.

[97] Haibin Liu, Vlado Keselj, Christian Blouin, and Karin Verspoor. Subgraph matching-based literature mining for biomedical relations and events. In *Proceedings of the 2012 AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, pages 32–37, 2012.

[98] R.E. Longacre. *The Grammar of Discourse*. NATO Advanced Study Institute Series. Springer Dordrecht, 1983.

[99] Zhiyong Lu. PubMed and beyond: A survey of web tools for searching biomedical literature. *Database*, 2011, 2011.

[100] J. Lyons and W. Lyons. *Semantics*. Language Arts & Disciplines. Cambridge University Press, 1977.

[101] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[102] Fiona Mao, Robert Mercer, and Lu Xiao. Extracting imperatives from wikipedia article for deletion discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 106–107, 2014.

[103] Brendan Marshall, Derin Benerci Keskin, and Andrew L. Mellor. Regulation of prostaglandin synthesis and cell adhesion by a tryptophan catabolizing enzyme. *BMC Biochemistry*, 2:5, 2001.

[104] Claudio Masolo, Alessander Botti Benevides, and Daniele Porello. The interplay between models and observations. *Applied Ontology*, 13(1):41–71, 2018.

[105] Deborah L. McGuinness and Frank van Harmelen. OWL web ontology language overview. W3C Recommendation, World Wide Web Consortium, February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[106] Mary L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

[107] Marvin Minsky. A framework for representing knowledge. Artificial Intelligence Memo No. 306, Massachusetts Institute of Technology A.I. Laboratory, 1974.

[108] Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487, 2006.

[109] Raquel Mochales and Marie-Francine Moens. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, 2011.

[110] Bruce Morgan, Daria Ezeriņa, Theresa N.E. Amoako, Jan Riemer, Matthias Seedorf, and Tobias P. Dick. Multiple glutathione disulfide removal pathways mediate cytosolic redox homeostasis. *Nature Chemical Biology*, 9(2):119–125, 2013.

[111] Huy Nguyen and Diane Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, 2015.

[112] Timothy J. Norman, Daniela V. Carbogim, Erik C.W. Krabbe, and Douglas Walton. Argument and multi-agent systems. In *Argumentation Machines*, pages 15–54. Springer, 2003.

[113] Nathan Ong, Diane Litman, and Alexandra Brusilovsky. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, 2014.

[114] Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 116–126, 2015.

[115] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[116] Joonsuk Park and Claire Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, 2014.

[117] Joonsuk Park, Arzoo Katiyar, and Bishan Yang. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44, 2015.

[118] Andreas Peldszus. Towards segment-based recognition of argumentation structure in short texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 88–97, 2014.

[119] Andreas Peldszus and Manfred Stede. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109, 2015.

[120] C. Perelman and L. Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, 1973.

[121] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, pages 1–40, 2012.

[122] Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2), 2008.

[123] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Natural Language Processing Using Very Large Corpora*, pages 157–176. Springer, 1999.

[124] K.E. Ravikumar, Haibin Liu, Judith D. Cohn, Michael E. Wall, and Karin Verspoor. Literature mining of protein-residue associations with graph rules learned through distant supervision. *Journal of Biomedical Semantics*, 3(3):S2, 2012.

[125] Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology*, 8(01):163–179, 2010.

[126] Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. A computational approach for generating Toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55, 2015.

[127] Hajo Rijgersberg, Mark Van Assem, and Jan Top. Ontology of units of measure and related concepts. *Semantic Web*, 4(1):3–13, 2013.

[128] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, (AAAI-96), pages 1044–1049, 1996.

[129] Martin Ringwald, Janan T. Eppig, James A. Kadin, and Joel E. Richardson. GXD: A Gene Expression Database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Research*, 28(1):115–119, 2000.

[130] William R. Roberts. *The Works of Aristotle, Vol. 11: Rhetorica*. Oxford: Clarendon Press, 1924.

[131] Cornelius Rosse and José L.V. Mejino Jr. A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, 2003.

[132] Glenn Rowe and Chris Reed. Argument diagramming: The Araucaria project. *Knowledge Cartography*, pages 163–181, 2008.

[133] Joseph Sambrook and David W. Russell. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 2001.

[134] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, 2015.

[135] Craig Schlenoff, Craig Schlenoff, Florence Tissot, John Valois, and Jintae Lee. The Process Specification Language (psl) Overview and Version 1.0 Specification. NISTIR 6459, National Institute of Standards and Technology, 2000.

[136] Jodi Schneider. Automated argumentation mining to the rescue? Envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argumentation Mining*, pages 59–63, 2014.

[137] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

[138] Stefan Schulz and Ludger Jansen. Formal ontologies in biomedical knowledge representation. *Yearbook of Medical Informatics*, 8(1):132–146, 2013.

[139] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[140] Isabel Segura-Bedmar, Paloma Martinez, and César de Pablo-Sánchez. Extracting drug-drug interactions from biomedical texts. *BMC Bioinformatics*, 11(Suppl5):P9, 2010.

[141] Rob Shearer, Boris Motik, and Ian Horrocks. HermiT: A highly-efficient OWL reasoner. In *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions*, (OWLED 2008) CEUR Workshop Proceedings Vol. 432, 2008.

[142] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

[143] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, 2015.

[144] Larisa Soldatova, Ross King, Piyali Basu, Emma Haddi, and Nigel Saunders. The representation of biomedical protocols. *EMBnet.journal*, 19(B), 2013.

[145] Larisa N. Soldatova and Ross D. King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11), 2006.

[146] Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, 2014.

[147] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[148] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, (EMNLP), pages 46–56, 2014.

[149] John Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[150] Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41, 2015.

[151] Simone Teufel. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. Center for the Study of Language and Information, 2010.

[152] Simone Teufel. Scientific argumentation detection as limited-domain intention recognition. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, CEUR Workshop Proceedings Vol. 1341, 2015.

[153] Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, 1999.

[154] Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[155] Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (EMNLP 2009), pages 1493–1502, 2009.

[156] Dorothea K. Thompson. Arguing for experimental "facts" in science: A study of research article Results sections in biochemistry. *Written Communication*, 10(1):106–128, 1993.

[157] Paul Thompson, Philip Cotter, Sophia Ananiadou, John McNaught, Simonetta Montemagni, Andrea Trabucco, and Giulia Venturi. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2159–2166, 2008.

[158] Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393, 2011.

[159] Christopher W. Tindale. *Acts of Arguing: A Rhetorical Model of Argument*. SUNY Press, 1999.

[160] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[161] Bianka Trevisan, Eva Dickmeis, Eva-Maria Jakobs, and Thomas Niehr. Indicators of argument-conclusion relationships. An approach for argumentation mining in german discourses. In *Proceedings of the First Workshop on Argumentation Mining*, pages 104–105, 2014.

[162] Richard Tzong-Han Tsai, Wen-Chi Chou, Yu-Chun Lin, Cheng-Lung Sung, Wei Ku, Ying-Shan Su, Ting-Yi Sung, and Wen-Lian Hsu. BIOSMILE: Adapting semantic role labeling for biomedical verbs: An exponential model coupled with automatically generated template features. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, (LNLBioNLP'06), pages 57–64, 2006.

[163] Richard Tzong-Han Tsai, Wen-Chi Chou, Ying-Shan Su, Yu-Chun Lin, Cheng-Lung Sung, Hong-Jie Dai, Irene Tzu-Hsuan Yeh, Wei Ku, Ting-Yi Sung, and Wen-Lian Hsu. BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. *BMC Bioinformatics*, 8:325, 2007.

[164] Tzong-Han Tsai, Chia-Wei Wu, Yu-Chun Lin, and Wen-Lian Hsu. Exploiting full parsing information to label semantic roles using an ensemble of ME and SVM via integer linear programming. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, (CONLL'05), pages 233–236, 2005.

[165] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*, pages 382–392. Springer, 2005.

[166] Frans H. van Eemeren and Rob Grootendorst. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press, 2004.

[167] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 2005.

[168] Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, (LREC-2012), pages 812–817, 2012.

[169] Vern Walker, Karina Vazirova, and Cass Sanford. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 1–10, 2014.

[170] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[171] Douglas Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, 1996.

[172] Adrienne M. Wang, Yoshinari Miyata, Susan Klinedinst, Hwei-Ming Peng, Jason P. Chua, Tomoko Komiyama, Xiaokai Li, Yoshihiro Morishima, Diane E. Merry, William B. Pratt, Yoichi Osawa, Catherine A. Collins, Jason E. Gestwicki, and Andrew P. Lieberman. Activation of Hsp70 reduces neurotoxicity by promoting polyglutamine protein degradation. *Nature Chemical Biology*, 9(2):112–118, 2013.

[173] Tuangthong Wattarujeekrit, Parantu K. Shah, and Nigel Collier. PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5:155, 2004.

[174] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. Ontonotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[175] Adam Wyner, Wim Peters, and David Price. Argument discovery and extraction with the Argument Workbench. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 78–83, 2015.

[176] Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument - Proceedings of COMMA 2012*, pages 43–50, 2012.

[177] Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. Learning sentence ordering for opinion generation of debate. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 94–103, 2015.

[178] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. Highway long short-term memory RNNs for distant speech recognition. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, (ICASSP), pages 5755–5759, 2016.

[179] Jie Zheng, Elisabetta Manduchi, and Christian J. Stoeckert Jr. Development of an application ontology for beta cell genomics based on the ontology for biomedical investigations. In *Proceedings of the 4th International Conference on Biomedical Ontology*, (ICBO 2013) CEUR Workshop Proceedings Vol. 1060, pages 62–67, 2013.

[180] Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, 2015.

APPENDICES

Appendix A

Annotation Guidelines, Questions, Observations, and Meeting Notes

Annotation Guidelines for Experimental Procedures

Version

April 14th, 2018

1- Introduction and background information

What is rhetorical move?

A rhetorical move can be defined as a text fragment that conveys a distinct communicative goal, in other words, a sentence that implies an author's specific purpose to readers.

What are the types of rhetorical moves?

There are several types of rhetorical moves. However, we are interested in 4 rhetorical moves that are common in the method section of a scientific article that follows the Introduction Methods Results and Discussion (IMRaD) structure.

1- **Description of a method**: It is concerned with a sentence(s) that describes experimental events (e.g., "Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008)).
2- **Appeal to authority**: It is concerned with a sentence(s) that discusses the use of standard methods, protocols, and procedures. There are two types of this move:
    - A reference to a well-established "standard" method (e.g., the use of a method like "PCR" or "electrophoresis").
    - A reference to a method that was previously described in the literature (e.g., "Protein was determined using fluorescamine assay [41]." (Larsen, Frandesn and Treiman, 2001)).
3- **Source of materials**: It is concerned with a sentence(s) that lists the source of biological materials that are used in the experiment (e.g., "All microalgal strains used in this study are available at the Elizabeth Aidar Microalgae Culture Collection, Department of Marine Biology, Federal Fluminense University, Brazil." (Larsen, Frandesn and Treiman, 2001)).
4- **Background information**: It is concerned with a sentence(s) that deals with method justifications, comments, or observations (e.g., "Unfortunately, our attempts to detect activation of S1P4 expressed in these cells … were unsuccessful. Therefore, we used CHOK1cells as an alternative host in these studies..." (Holdsworth et al. 2004).

What is a semantic role?

A semantic role is "the underlying relationship that a participant has with the main verb in a clause"[1]. For example,

[1] Semantic Role. (2015, December 3). Retrieved August 17, 2017, from http://www.glossary.sil.org/term/semantic-role

"An apple was eaten by John"

The sentence describes a frame "eating an apple", so "John" is the experiencer or agent who eats the apple, and the object "apple" is the patient which is being eaten.

What are the kinds of semantic roles?

There are various semantic roles which already were developed in the literature (e.g., Verb Net).

1- **Predicate**: The verb that initiates the frame. It could be a verb or a nominalized verb. Basically, nominalization is "to convert (another part of speech) into a noun, as in changing the adjective *low* into *the lowly* or the verb *legalize* into *legalization*"[2].
2- **Agent**: Initiator of action, capable of volition and most of the times the agent come in phrase like "we" or "the authors" ➔ Proteins were washed three times by **the authors**.
3- **Patient**: Affected by action, undergoes change of state ➔ **Haplotypes of the individuals** were reconstructed using BEAGLE (version 4.0) (Browning and Browning 2007).
4- **Theme**: Not changed by an action, or being "located" ➔ **Other computing works of this report** were conducted in R (version 2.14.2) (R Core Team 2015), a free software environment for statistical computing and graphics.
5- **Instrument**: "used for objects (or forces) that come in contact with an object and cause some change in them. Generally introduced by 'with' prepositional phrase"[3]. Most of the time appears as a Prepositional Phrase (PP).

   We have created sub-categories under the semantic role "Instrument" to include:

   Note: bold-faced words or phrases are the ones that are referred to in each example.

   - (Change) a thing or protocol that can change another thing(s).
     - Note that if the sentence describes selective media which allow only the selected cells to survive while others not. In this case, we should label it as instrument of change.

     Example: "Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) **with pulldown buffer** and proteins harvested **in SDS-sample buffer**, separated **by SDS-PAGE**, and analyzed by autoradiography." (Ester & Uetz, 2008).

---

[2] http://www.dictionary.com/browse/nominalize

[3] Martha Palmer | Projects | Verb Net. (n.d.). Retrieved August 21, 2017, from http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

- (Measure) a thing or protocol that can measure another thing(s).
  Example: "Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed **by autoradiography**." (Ester & Uetz, 2008).
- (Observe) a thing which can be used to observe another thing(s)
  Example:" The mitochondria was observed **by spinning disk confocal microscopy**".
- (Maintain) a thing or protocol which can be used to maintain the state of another thing(s).
  You should note that:
  - If the sentence contain inhibition process, you should label it as maintain.
  - If the sentence describe a media that used for growth, you should label it as maintain too.

  Example: "Once the samples were in EPR tubes, they were immediately frozen in liquid nitrogen, and stored **in liquid nitrogen** before using." (Chen & Guidotti, 2001).
- (Catalyst) a thing that can be used as a catalytic "facilitator" (there are two different types of enzymes, one needs a cofactor to be active as a catalyst and the other doesn't need a cofactor).
  Example: "The ca. 900 bp PCR products were digested **with NdeI and HindIII** and ligated into pUC19." (Carenbauer et al., 2002)
- (Mathematical) a mathematical or computational instrument (e.g., simulation, algorithm, equation and the use of software)
  Example:" Simulations of these EPR spectra were accomplished **with the computer program QPOWA** [30,31])." (Chen & Guidotti, 2001)
- (Reference) a reference to a paper that describes the complete protocol.
  Example: "The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase were **as previously described [11]**." (Brown et al., 2000)

Other types of semantic roles that occur in some frames:

6- **Goal:** We have categorized into two types:

a- Physical: A thing that already existed and an action is directed toward it or place to which something moves. ➔ "The ca. 900 bp PCR products were digested with NdeI and HindIII and ligated **into pUC19**."

b- Purpose: Used to state author's intention for doing something. ➔ "**To monitor luciferase cycling**, 0.122 Ã— 10 6 cells were seeded per 35-mm plate."

7- **Location**: The physical place where the experiments took place. ➔ "The DNA sequences were analyzed by **the Biosynthesis and Sequencing Facility in the Department of Biological Chemistry at Johns Hopkins University**."
8- **Factitive:** it comes into existence as a result of the event. ➔ **Plasmid libraries** were generated through a two-step cloning process (Kwasnieski et al. 2012 , 2014 ; White et al. 2013).
9- Protocol detail:
    1- **Temperature**: usually comes after the instrument as a prepositional phrase that states the process temperature.
    Example: "Beads with bound proteins were washed six times (for 10 min under rotation **at 4°C**) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008).
    2- **Time (duration)**: "class-specific role that is used to express time" (Verb Net project). Usually comes after the instrument as a prepositional phrase that states the process time.
    Example: "Beads with bound proteins were washed six times (**for 10 min** under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008).
    3- **Repetition** of a process:
    Example: "Beads with bound proteins were washed **six times** (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008).
    4- **Condition** of a process or the manner in which it was carried out:
    Example: "Beads with bound proteins were washed six times (for 10 min **under rotation** at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008).
    5- **Cofactor:** is classified as "inorganic substances that are required for, or increase the rate of, catalysis."[4] Please see the list of most common buffers in Appendix A.
    Example:" For phosphorylation, three identical reactions contained H5R protein (70 pmol), B1R protein kinase (90 µl), Tris-HCl, pH 7.4 (20 mM), **magnesium chloride (5 mM), ATP (50 µM), [γ-32P] ATP (50 µCi) and dithiothreitol (2 mM)** in a total volume of 500 µl." (Brown et al., 2000)
    6- **Coenzyme:** is defined as "an organic molecule that is required by certain enzymes to carry out catalysis."[4] (In this study, we will call both coenzyme and cofactor a "cofactor". Cofactor is a hypernym). (Needs to be embedded with the definition above.

4 coenzymes and cofactors. (n.d.). Retrieved September 23, 2017, from http://academic.brooklyn.cuny.edu/biology/bio4fv/page/coenzy_.htm

7- **Buffer:** is defined as "a solution containing either a weak acid and a conjugate base or a weak base and a conjugate acid, used to stabilize the pH of a liquid upon dilution."[5] Please see the list of most common buffers in Appendix B.
Example:" For phosphorylation, three identical reactions contained H5R protein (70 pmol), B1R protein kinase (90 µl), **Tris-HCl, pH 7.4 (20 mM),** magnesium chloride (5 mM), ATP (50 µM), [γ-32P] ATP (50 µCi) and dithiothreitol (2 mM) in a total volume of 500 µl." (Brown et al., 2000)

- Examples of annotating semantic roles from our dataset

a- "Beads with bound proteins were washed six times (for 10 min under rotation at 4°C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography." (Ester & Uetz, 2008).

Event 1	Event 2	Event 3	Event 4
Patient: Beads with bound proteins Predicate: were washed Instrument (change): Pulldown buffer Protocol detail: - Repetition: six times - Time: 10 min - Condition: under rotation - Temp: 4 C	Patient: Proteins Predicate: harvested in Instrument (change): SDS-sample buffer	Patient: Proteins Predicate: separated by Instrument (change): SDS-PAGE	Patient: Proteins Predicate: analyzed by Instrument (measure): Autoradiography

b- "The ca. 900 bp PCR products were digested with NdeI and HindIII and ligated into pUC19." (Carenbauer et al., 2002)

5 Buffer - Biology-Online Dictionary. (n.d.). Retrieved September 23, 2017, from http://www.biology-online.org/dictionary/Buffer

Event 1	Event 2
Patient: The ca. 900 bp PCR products Predicate: were digested Instrument (catalyst): with NdeI and HindIII	Patient: The ca. 900 bp PCR products Predicate: ligated Goal: into pUC19

c- "The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase were as previously described [11]." (Brown et al., 2000)

Event 1
Patient: The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase Instrument (reference): [11] Predicate: described

d- "A large peak of radioactivity (unreacted ATP) eluted with the water and was discarded, and a smaller broad peak of radioactivity that eluted with 50% acetonitrile was retained and concentrated to 200 µl by rotary evaporation." (Brown et al., 2000)

Event 1
Patient: a large peak of radioactivity (unreacted ATP) which was already eluted with the water Predicate: discarded
Event 2 Patient: a smaller broad peak of radioactivity which was already eluted with 50% acetonitrile Predicate: retained **Event 3** Patient: a smaller broad peak of radioactivity which was already eluted with 50% acetonitrile Predicate: concentrated Instrument (change): rotary evaporator

e- “Peptides were sequenced on an Applied Biosystems 476A protein sequencer and phosphorylation sites were analyzed using solid phase Edmann sequencing [20].” (Brown et al., 2000)

Event 1
Patient: peptides Predicate: sequenced Instrument (measure): an Applied Biosystems 476A protein sequencer Event 2 Patient: Phosphorylation sites Predicate: analyzed Instrument (mesure): using solid phase Edmann sequencing [20]

f- “Steady-state kinetics constants, Km and kcat, were determined by fitting initial velocity versus substrate concentration data directly to the Michaelis equation using CurveFit [36]. ” (Carenbauer et al., 2002)

Event 1
Patient: Steady-state kinetics constants, Km and kcat Predicate: determined Instrument (mathematical): by fitting initial velocity versus substrate concentration data directly to the Michaelis equation using CurveFit [36]

2- Annotation guidelines

These guidelines describe a classification scheme for the Method section in biochemistry articles which are concerned with the rhetorical moves and semantic roles.

1- Before the annotation

The annotator should read the entire article. This is very important as we are only looking for the annotation of one section in biochemistry articles (i.e., the Method section). Thus, the interpretation of some sentences in the Method section can become clear once the entire article has been read. Please note that you don't need to understand the article in detail, you can go back and forth between sections in the article. Please also try to focus on the main four rhetorical moves and ensure that the sentence is concerned with one of these moves.

2- During annotation

Annotation should be proceed by only annotating one sentence at a time and assigning this sentence to one of the moves. Usually consecutive sentences are marked with the same move type. You can label consecutive sentences with the same move if these sentences share the same move. First you should find the verbs in every sentence, and then the annotator should be able to answer the following questions:

Q1. Can you identify the predicate (e.g., verb, nominalized verb, adjective) in the sentence?

If yes, you should label it as "predicate". Then go to Q2

Q2. Can you identify the patient, theme, or factitive from this sentence or phrase? If yes, you should label it as either patient, theme or factitive depends on their definition giving in (section 1), and then proceed to Q3.

Q3. Can you identify the instrument in this sentence or phrase?

If yes, you should label it as "instrument" and select one of the instrument types (e.g., Change, Measure, Observe, Maintain, Catalyst, Mathematical and Reference) then proceed to Q4. If not, you can proceed to Q4

Q4. Is there additional information in the sentence such as process temperature, time, or buffer used in the experiment?

If yes, you should list this information under protocol detail and proceed to Q5. If no, you should proceed to Q5.

Q5. Can you identify a goal either (**physical**: where the theme, patient, or factitive in this sentence is directed to OR **purpose**: which indicates the author intention in the sentence?

If yes, you should label it as "Goal: physical" for the first type or as "Goal: purpose" for the second one, and proceed to Q6. If no, you should proceed to Q6.

Q6. Can you identify the location where the experimental process took place?

If yes, you should label it as “location” and then proceed to Q7. If no, you should proceed to Q7.

Q7. Does this sentence describe an experimental procedure?

If yes, you should label it as “description of the method” and proceed to next sentence in the paragraph. If no, you should proceed to Q8.

Q8. Does this sentence use a technique, protocol or method that was previously introduced in the scientific field?

If yes, you should label it as “appeal to authority” and proceed to next sentence in the paragraph. If no, you should proceed to Q9.

Q9. Does this sentence talk about method justifications, comments, or observations?

If yes, you should label it with “background information” and proceed to next sentence in the paragraph. If no, you should proceed to Q10.

Q10. Does this sentence list or describe experimental materials?

If yes, you should label it as “source of the materials” and proceed to examine the next sentence. If no, you should proceed to next sentence in the paragraph.

Appendix A:

A List of most common buffers[6]:

Buffer
MES
Bis-Tris
ADA
ACES
PIPES
MOPSO
Bis-Tris Propane
BES
MOPS
TES
HEPES
DIPSO
MOBS
TAPSO
Trizma
HEPPSO
POPSO
TEA
EPPS
Tricine
Gly-Gly
Bicine
HEPBS
TAPS
AMPD
TABS
AMPSO
CHES
CAPSO
AMP
CAPS
CABS

[6] http://www.sigmaaldrich.com/life-science/core-bioreagents/biological-buffers/learning-center/buffer-reference-center.html

Appendix B:

A List of most common cofactors[7]:

Cofactor
Thiamine pyrophosphate [29]
NAD^+ and $NADP^+$ [30]
Pyridoxal phosphate [31]
Methylcobalamin [32]
Cobalamine [5]
Biotin [33]
Coenzyme A [34]
Tetrahydrofolic acid [35]
Menaquinone [36]
Ascorbic acid [37]
Flavin mononucleotide [38]
Flavin adenine dinucleotide [38]
Coenzyme F420 [39]
Adenosine triphosphate [40]
S-Adenosyl methionine [41]
Coenzyme B [42]
Coenzyme M [43][44]
Coenzyme Q [45]
Cytidine triphosphate [46]
Glutathione [47][48]
Heme [49]
Lipoamide [5]
Methanofuran [50]
Molybdopterin [51][52]
Nucleotide sugars [53]
3'-Phosphoadenosine-5'-phosphosulfate [54]
Pyrroloquinoline quinone [55]
Tetrahydrobiopterin [56]
Tetrahydromethanopterin [57]

[7] https://en.wikipedia.org/wiki/Cofactor_%28biochemistry%29

Annotation Q/A & Observations

22 Jan 2018 / 7:10 PM / UW - DC

Contributors

Mohammed Alliheedi, Bob Mercer, Sandor Hass-Neil, Beth Locke, Danial Mohsin, and Silvia Gan

Q/A

Here is a list of questions/answers regarding labeling sentences in the method section of biochemistry articles.

Q1- Can we consider 35-mm plate and 15-cm plate in the following examples as instrument or goal? “To monitor luciferase cycling, 0.122 Ã— 10 6 cells were seeded per **35-mm plate**.For the purposes of collection, 2.25 Ã— 10 6 cells were seeded per **15-cm plate**.”
In those sentences I would say the plates are more of a goal but in a different context I suppose you could think of them as maintaining instruments (for cell growth conditions).

Q2- Would cell media be considered instruments of maintenance or growth?
The media maintains optimal conditions for growth but it is the cells that grow themselves. So, we should mark it as a maintaining instrument.

Q3- Should we label all media as instruments of maintenance?
No, not all of the times. Selective media should be labeled as instrument of change because this type of media allow only the selected cells to survive while others not.

Q4- The following example was taking from the method section of an article which refer to other section “supplemental methods” in the same article. if they references their Supplemental methods, is this considered "Description of methods" or "Reference to methods"? Example: "Unique vervet sequences represent the best source for detecting new lineage-specific genomic elements, including active retrotransposons, and were extracted as described in **Supplemental Methods**."
Within the context of the sentence I would call that a reference to methods. Just as they can reference to their own/another paper, a reference to supplemental methods is asking you to look at something other than what you are currently reading in order to find out what was actually done.

Q5- How do you annotate in-text citations?

You can label the span of in-text citations as in the following example:
“For all binding sites, we utilized 120-bp of sequence for our assays centered on the CEBPB ChIP-seq binding site summits as determined by **MACS (Zhang et al. 2008)**”

Comments/Observations

1- The growth/selective media example:

“Saccharomyces cerevisiae yeast strains used in this study listed in Supplemental Table 2 were derived from W303, grown in rich (YPAD) or selective media at 30°C”

---> Here, *the media* is an instrument *of change* rather than instrument of maintenance since the operator is selecting the W303 derived yeast cells only thus is causing a change in the system.

But if, for example, the next step was to grow the selected cells in an optimum growth media, this (optimum) media should be annotated as instrument *of maintenance*, since this media is just allowing the cells to divide and naturally grow.

Notes

- **This list will be updated periodically.**
- **Everyone is encouraged to ask questions or suggest comments about the annotation.**

Meeting Notes – May 18, 2018

Example 1:

Beads with bound protein [Patient, not theme] were washed….

Example 2:

"Flow cytometry measurements on the strain GFP-HHF1 were used to define gates on FSC-A/FSC -O and FSC-A/SSC-A that best separate G1 from G2 cells."

- 2 ways to interpret this sentence – both interpretations are shown below but we decided that the **SECOND INTERPRETATION** is the most correct since it properly labels more of the required details a researcher may need

1. **Predicate:** *were used*
   Theme: *Flow cytometry measurements on the strain GFP-HHF1*
   Goal: *used to define gates on FSC-A/FSC -O and FSC-A/SSC-A that best separate G1 from G2 cells.*

2. **Predicate:** *to define*
   Theme: *Flow cytometry measurements on the strain GFP-HHF1*
   Factitive: *gates on FSC-A/FSC -O and FSC-A/SSC-A*
   Condition: *that best separate G1 from G2 cells.*

Discussing location & companies:

- Phrases like "from Glasgow" or "from Germany" can be labelled as **Location**
- Bracketed references to a company from which a material is sourced in the description of a method should be included in the theme

  i.e. As a negative control, **an equal amount of AllStars Negative Control siRNA (Qiagen) [Theme]** ***was used for transfection [predicate]*** in parallel.

Discussing nominalized verbs and ArgMoves:

i.e. Cells were then synchronized with dexamethasone following transfection as described above.

Theme: Cells
Predicate1/Verb: were then synchronized
Instrument/Maintain: with dexamethasone
Predicate2/Nominalized Verb: transfection
Instrument/Reference: as described above

ArgMove1/Desc of Method: Cells were then synchronized with dexamethasone
ArgMove2/Ref to Method: following transfection as described above

Discussing Goal/Physical vs Condition:

i.e 1.5 x 10^4 cells were seeded per 15 mm plate.

- The number of cells and the size of the plate are important to researchers
- They are used as measures of the concentration without the authors specifically stating the concentration as a value like 10 g/mL
- Label both parts as condition as follows:

Theme: cells
Predicate: were seeded
Condition1: 1.5 x 10^4 cells
Condition2: per 15 mm plate
Goal/Physical: plate

Meeting Notes – May 18, 2018

Example 1:

Beads with bound protein [Patient, not theme] were washed….

Example 2:

"Flow cytometry measurements on the strain GFP-HHF1 were used to define gates on FSC-A/FSC -O and FSC-A/SSC-A that best separate G1 from G2 cells."

- 2 ways to interpret this sentence – both interpretations are shown below but we decided that the **SECOND INTERPRETATION** is the most correct since it properly labels more of the required details a researcher may need

1. **Predicate:** *were used*
   Theme: *Flow cytometry measurements on the strain GFP-HHF1*
   Goal: *used to define gates on FSC-A/FSC -O and FSC-A/SSC-A that best separate G1 from G2 cells.*

2. **Predicate:** *to define* *>> USED TO DEFINE*
   Theme: *Flow cytometry measurements on the strain GFP-HHF1*
   Factitive: *gates on FSC-A/FSC -O and FSC-A/SSC-A*
   Condition: *that best separate G1 from G2 cells.*

Discussing location & companies:

- Phrases like "from Glasgow" or "from Germany" can be labelled as **Location**
- Bracketed references to a company from which a material is sourced in the description of a method should be included in the theme

  i.e. As a negative control, **an equal amount of AllStars Negative Control siRNA (Qiagen) [Theme]** ***was used for transfection [predicate]*** in parallel.

Discussing nominalized verbs and ArgMoves:

i.e. Cells were then synchronized with dexamethasone following transfection as described above.

~~**Theme:**~~ **Patient:** Cells
Predicate1/Verb: were then synchronized with
Instrument/Maintain: dexamethasone
Predicate2/Nominalized Verb: transfection
Instrument/Reference: as described above

ArgMove1/Desc of Method: Cells were then synchronized with dexamethasone
ArgMove2/Ref to Method: following transfection as described above

Discussing Goal/Physical vs Condition:

i.e 1.5 x 10^4 cells were seeded per 15 mm plate.

- The number of cells and the size of the plate are important to researchers
- They are used as measures of the concentration without the authors specifically stating the concentration as a value like 10 g/mL
- Label both parts as condition as follows:

Theme: cells
Predicate: were seeded
Condition1: 1.5 x 10^4 cells
Condition2: per 15 mm plate
Goal/Physical: plate

Meeting notes and annotation examples (May 19, 2018)

Flow cytometry measurements on the strain GFP-HHFS were used to define gates on FSC-A/FSC-0 and FSC-A/SSC-A that best separate G1 from G2 cells.

Predicate: were used to define
Theme Flow cyt……GFP-HHFS
Factitive: gates on…SSC-A
Condition: That best separate….g2 cells

a) *ATP, ADP, NDP, PNP were purchased from Sigma (st louis, Mo).*

Predicate: were purchased from
Theme ATP-PNP
Location: Sigma (st louis Mo)
Arg Move: Source materials.

b) *growth factor (Sigma-Aldrich)*

Protocol details: Co factor: Growth factor (Sigma Aldrich)

c) *1.5 x 10^4 cells were seeded per 15 mm plate.*

Theme: cells

Predicate: were seeded

Condition 1: 1.5 x 10^4 cells

Condition 2: per 15 mm plate

Goal/Physical: plate

d) *The accession numbers are detailed in Supplemental Table S1.*
Predicate: are detailed in
Instrument_ref: Supplemental Table S1
Arg.move: background info

Data on histone marks and transcription factors in mESCs were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/).
Predicate: were downloaded from
Instrument_ref: the GEO database (http://www.ncbi.nlm.nih.gov/geo/).
Arg.Move: Background info.

e) *Next-generation sequencing library preparation was performed as previously described (Kwasnieski et al. 2012 , 2014 ; White et al. 2013).*

Predicate: Was performed
Instrument_ref: as previously described (Kwasnieski et al. 2012 , 2014 ; White et al. 2013).
Arg.Move: Ref_method

f) *The findings of this study have been submitted to the GEO database.*

Predicate: Have been submitted to
SemanticRole_Goal: GEO database.
Arg.Move: background info.

g) *The data was compared with the NCBI database.*

Predicate: was compared with.
Instrument_ref: NCBI database
Arg.move: description method

h) *As a negative control, an equal amount of siRNA (Qiagen) was used for transfection in parallel.*

Predicate: was used for transfection
Patient: siRNA(Qiagen)
Goal: As a negative control
Condition 1: in parallel
condition 2: an equal amount of siRNA(Qiagen)

i) Cells were then synchronized with dexamethasone following transfection as described above.

Patient: Cells (id with predicate 1 and predicate 2)
Predicate1/Verb: were then synchronized with
Instrument/Change: dexamethasone
Predicate2/Nominalized Verb: transfection
Instrument/Reference: as described above (id with predicate 2)

j) RNA from the cells was extracted using the RNA miniprep kit (Qiagen)

Predicate: was extracted using
Inst_Change: RNA miniprep kit(Qiagen)
Theme: RNA from the cells
ARg.Move: Description Method

Appendix B

Annotation for Experimental Procedures with Domain Expert

M1: Description of method; M2: Reference to a protocol; M3: Source of materials; M4: Background information

Processes:

1- The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase were as previously described [11].
2- In some experiments a trpE-H5R fusion protein (pATH11-Ag35) was used [19]. (Alternative to previous protein)
3- For phosphorylation, three identical reactions contained H5R protein (70 pmol), B1R protein kinase (90 µl), Tris-HCl, pH 7.4 (20 mM), magnesium chloride (5 mM), ATP (50 µM), [γ-32P] ATP (50 µCi) and dithiothreitol (2 mM) in a total volume of 500 µl.
4- Incubation was for 30 min at 30° C.
5- The 500 µl reaction mixtures, above, were adjusted to 50 mM Tris-HCl, pH 7.4, and 0.01% reduced Triton X-100 at a final volume of 600 µl.
6- To this was added 0.4 µg V8 protease (Boehringer Mannheim) and incubation carried out at 30° C for 18 h.
7- To prepare the peptides for HPLC analysis the reaction mixtures were pooled and applied to a SEP-PAK cartridge (prewashed with successive 10 ml portions of 50% acetonitrile and water) and eluted with water (40 ml) followed by 50% acetonitrile (40 ml) and finally 100% acetonitrile (30 ml).
8- A large peak of radioactivity (unreacted ATP) eluted with the water and was discarded, and a smaller broad peak of radioactivity that eluted with 50% acetonitrile was retained and concentrated to 200 µl by rotary evaporation.

Event 1	Event 2	Event 3	Event 4
M2 Theme: The preparation of authentic vaccinia H5R protein and recombinant B1R protein kinase Instrument (reference): using method in [11] Predicate: (prepared) Event_Type: Data collection Notes: we can create an event for sentence 2 (trpE-H5R fusion protein) which can be essential for second tier of information.	M1 Theme: H5R protein (70 pmol) Predicate: (performed) (in this sense means phosphorylated) Instrument (change): B1R protein kinase (90 µl), Tris-HCl, pH 7.4 (20 mM), magnesium chloride (5 mM), ATP (50 µM), [γ-32P]ATP (50 µCi) and dithiothreitol (2 mM) in a total volume of 500 µl Instrument (maintain) : incubator Comment: there are three replicates of the above instrument (change). Protocol detail (instrument-maintain) Buffer: Tris-HCl, pH 7.4 (20 mM) (instrument-catalyst) Cofactor: magnesium chloride (5 mM), ATP (50 µM), [γ-32P]ATP (50 µCi) dithiothreitol (2 mM) Temp: at 30° C Time: (for 30 min) Event_Type: Data collection	M1 Theme: The 500 µl reaction mixtures Instrument (change): pipetting (from the ontology which is being implied) (100 µl) Tris-HCl and reduced Trition X-100 Predicate: adjusted Event_Type: Data collection	M1 Theme: The mixture of 600 µl Instrument (catalyst): V8 protease (Boehringer Mannheim) Instrument (maintain): incubator Instrument (maintain): (buffer) 50 mM Tris-HCl, pH 7.4 Predicate: added but the verb (Proteolysis – lysis of a protein) is used to describe why the addition has been done. (Implied knowledge from the domain, protease is enzyme that catalysis the cutting of protein). Protocol detail: Incubation Temp: at 30° C Incubation Duration: for 18 h Event_Type: Data collection

Event 6	Event 7	Event 8	Event 9
M1 Theme: the reaction mixtures Instrument: N/A Goal: a SEP-PAK cartridge Predicate: pooled & applied Result: To prepare the peptides for HPLC (high performance liquid chromatography) analysis Event_Type: Data collection	M1 Theme: the reaction mixtures Instrument (change): with water (40 ml) followed by 50% acetonitrile (40 ml) and finally 100% acetonitrile (30 ml) Predicate: eluted Result: To prepare the peptides for HPLC analysis Event_Type: Data collection	M1 Theme: a large peak of radioactivity (unreacted ATP) which was already eluted with the water Instrument: N/A Predicate: discarded Event_Type: Data collection (Notes: different things elute in different elution buffers, so the unreacted ATP elutes with water)	M1 Theme: a smaller broad peak of radioactivity which was already eluted with 50% acetonitrile instrument (change): rotary evaporator Predicate: retained (in this sense means collected from the domain) & concentrated Result: a solution of 200 µl that containing the protein that get phosphorylated. Event_Type: Data collection Note (a smaller broad peak of radioactivity is the protein that get phosphorylated) Note: there is information that would be knowing for any working biochemist which that the theme that we already described in event 9 is referred to the phosphorylated H5R protein. This type of information is implicit and expert in the field would infer that.

9- Initial purification was with a Vydac protein and Peptide C18 column (25 × 04 cm) on a Gilson HPLC system, and this was followed by further purification on a Vydac C18 2.1 × 180 mm microbore column.

10- Details of the gradients used are given in the text. (additional information)

11- Peptides were sequenced on an Applied Biosystems 476A protein sequencer and phosphorylation sites were analyzed using solid phase Edmann sequencing [20].

12- Synthetic peptides were purchased from Thistle Research, Glasgow, UK. (source of material) (additional information for second tier in order to replicate what the researchers have done in this paper)

13- Each peptide (3 mM) was incubated with [γ-32P]ATP (6.3 µCi) B1R protein kinase (4 µl), Tris-HCl, pH 7.4 (20 mM), magnesium chloride (5 mM), ATP (50 µM), and dithiothreitol (2 mM) in a total volume of 20 µl. Incubation was for 30 min at 30° C.

14- The reaction mixtures were applied in 1 cm strips to thin layer cellulose plates and subjected to electrophoresis for 4 h at 200 V in a solution of pyridine : acetic acid : water (20:200:1780) at pH 3.5.

15- The plates were dried, stained with ninhydrin to locate the unphosphorylated peptides, and subjected to autoradiography.

Event 10	Event 11	Event 12	Event 13
M1 Theme: H5R peptides (which is implied) Instrument (change): with a Vydac protein and Peptide C18 column (25 × 04 cm) on a Gilson HPLC system Predicate: purification Protocol detail: Vydac protein and Peptide C18 column (25 × 04 cm) on a Gilson HPLC system Event_Type: Data collection	M1 Theme: H5R peptides Instrument (change): on a Vydac C18 2.1 × 180 mm microbore column on a Gilson HPLC system Predicate: followed by further purification (purified) Event_Type: Data collection	M1 Theme: H5R (implied info) peptides Predicate: sequenced Instrument (measure): an Applied Biosystems 476A protein sequencer Event_Type: Data collection	M1 Theme: phosphorylation sites of (H5R peptides—implied info) Predicate: analyzed Instrument (measure): using solid phase Edmann sequencing Event_Type: Data analysis

Event 14	Event 15	Event 16	Event 17
M1 Theme: Each peptide (3 mM) instrument (change): with [γ-32P]ATP (6.3 µCi) B1R protein kinase (4 µl), Tris-HCl, pH 7.4 (20 mM), magnesium chloride (5 mM), ATP (50 µM), and dithiothreitol (2 mM) in a total volume of 20 µl Instrument (maintain): incubator Predicate: phosphorylation (implied info) Incubation Time: for 30 min Incubation Temp: at 30° C Event_Type: Data collection	M1 Theme: The reaction mixtures Predicate: applied Goal: cellulose plates Instrument: N/A Factitive: 1 cm strips (the description of how the mixture being applied in the plates.) Event_Type: Data collection	M1 Theme: The reaction mixtures Predicate: subjected Instrument (change): electrophoresis Protocol detail: Time: for 4 h Volt: at 200 V Buffer: in a solution of pyridine : acetic acid : water (20:200:1780) at pH 3.5. Event_Type: Data collection	M1 Theme: The plates Predicate: dried instrument: N/A Event_Type: Data collection

Event 24	Event 25
M1 Theme: The plates Predicate: stained Instrument (change): ninhydrin Result: to locate the unphosphorylated peptides Event_Type: Data collection	M1 Theme: The plates Predicate: subjected Instrument (measure): (autoradiograph) implied from the “to autoradiography” Event_Type: Data collection

Important steps in the paper:

1. Elution with water

2. Collecting flow through and replacing flow through container

3. Elution with 50% acetonitrile

4. Collecting flow through and replacing flow through container

5. Elution with 100% acetonitrile

6. Collecting flow through

- Digestion of proteins generate peptides.
- Important notes: protease and kinase are both enzymes that used to as catalyst for cutting or phosphorylation respectively. There are few other enzymes that do other functions.
- There is a list of enzymes and their functions here (https://en.wikipedia.org/wiki/List_of_enzymes).
- To use a semantic role which is linguistically found, a semantic role should be mentioned in the sentence.
- However, there are few roles that are not explicitly mentioned in the sentence but they can be inferred from the sentence.
- For event 10, we need to know the following information to fully understand the sentence.
- “Smaller things come out in the water including both isotopes of ATP, and in the 50% acetonitrile the phosphorylated protein is eluted”

M1: Description of method; M2: Reference to a protocol; M3: Source of materials; M4: Background information

Cell culture and preparation of soluble CD39

1- sCD39 transfected stable HighFive™ insect cells were cultured as described by Chen and Guidotti [25].
2- Soluble CD39 were purified as described [25] with some modifications.
3- After concanavalin A-Sepharose 4B and nickel affinity column chromatography, the ammonium sulfate precipitated sCD39 was collected and resuspended in about 50 µl of 40 mM Tris-HCl (pH7.5).
4- This sample was loaded on a Superose-12HR gel filtration column from Pharmacia Biotech equilibrated with 40 mM Tris-HCl (pH7.5).
5- The fractions containing the major peak were collected, and the solvent was changed to 20 mM Hepes (pH8.0), 120 mM NaCl, 5 mM KCl with an YM30 centricon from Millipore.
6- The final volume of the sample was around 200 µl, and the concentration of sCD39 was around 0.1 mM. (additional information for 2nd tier)
7- Concentrations of proteins were determined using DC Protein Assay from BIO-RAD using the provided protocol.

Event 1	Event 2	Event 3	Event 4
M2 Theme: sCD39 transfected stable HighFive™ insect cells Instrument (reference): as described by Chen and Guidotti [25] Predicate: (cultured) Event_Type: Data collection	M2 Theme: Soluble CD39 Predicate: purified Instrument (reference): as described [25] (with some modification) – for 2nd tier of information Part of event 2: Theme: sCD39 Instrument (change): concanavalin A-Sepharose 4B and nickel affinity column chromatography Predicate: purified (from the ontology – implied information) (NH4)2SO4 is the ammonium sulfate that is used to precipitate the sCD39 Event_Type: Data collection ---- additional info is in the last page too ---	M1 Theme: sCD39 Predicate: precipitated Instrument (change): the ammonium sulfate ---- comments------ This event wasn't Predicated by a verb and it was Predicated by adjectival phrase. Also it should be noted that this event was extracted from a phrase not a sentence.	M1 Theme: the ammonium sulfate precipitated sCD39 Instrument: unknown Predicate: collected Event_Type: Data collection The action terms for collection could be: Scrapping Melting Pipetting and so on

Event 5	Event 6	Event 7	Event 8
M1 Theme: the ammonium sulfate precipitated sCD39 Instrument (change): (the chemistry of this buffer or element) in about 50 µl of 40 mM Tris-HCl (pH7.5) Predicate: resuspended Goal: buffer Event_Type: Data collection	M1 Theme: the suspended sCD39 (The cue word here is "this sample" from the sentence) Goal: on a Superose-12HR gel filtration column from Pharmacia Biotech equilibrated with 40 mM Tris-HCl (pH7.5) Predicate: loaded --------comment---- Basically the sample is being entered or loaded into the column. Event_Type: Data collection	M1 Theme: The fractions containing the major peak Instrument (change & measure): on a Superose-12HR gel filtration column from Pharmacia Biotech equilibrated with 40 mM Tris-HCl (pH7.5) Predicate: collected Event_Type: Data collection	M1 Theme: the solvent – the buffer from the previous events containing the sCD39 The solute: sC39 Instrument (change): an YM30 centricon from Millipore Predicate: changed Changed to (we don't have a semantic role to label this kind of information): changed to 20 mM Hepes (pH8.0), 120 mM NaCl, 5 mM KCl Event_Type: Data collection

Nucleotidase activity assay and nucleotide separation by HPLC

8- The reactions (of Nucleotidase activity assay) were carried out in 20 mM HEPES-Tris (pH 7.0), 120 mM NaCl, and 5 mM KCl; they were started by adding nucleotides at 37°C. After incubation for 15 minutes, the reactions were stopped with 2% perchloroacetic acid.
9- Nucleotides were separated by HPLC on an anion exchange column (a 10 × 0.46 mm SAX column from Rainin Instruments) based on the method of Hartwick and Brown [29].
10- The low concentration buffer (A) was 0.08 M $NH_4H_2PO_4$ (pH3.8), and the high concentration buffer (B) was 0.25 M $NH_4H_2PO_4$ (pH4.95) with 8 mM KCl. (Additional information – Protocol detail)
11- The gradient used was 4 min, 0–2.5% (B); 26 min, 2.5–25% (B). Equilibration was done with buffer (A) for 10 minutes, and the flow rate was 1 ml/min. (Additional information – Protocol detail)

Event 9	Event 10	Event 11	Event 12
M1 Theme: Concentrations of proteins Instrument (measure): using DC Protein Assay from BIO-RAD using the provided protocol Predicate: determined ------comment--------- Assay means a protocol to elucidate new information	M1 Theme: The reactions (Nucleotidase activity assay) Instrument (maintain): in 20 mM HEPES-Tris (pH 7.0), 120 mM NaCl, and 5 mM KCl Predicate: carried out incubator: an instrument to maintain CD39: instrument that catalyses cleavage of a nucleoside (((CD39/ nucleotides) Buffer to maintain) Incubator) --- Temp: 37°C Time: 15 min Buffer and incubator are instruments to maintain the condition and CD39 as instrument to do the cleavage. Event_Type: Data collection	Sub-event for event 10 Additional information provided: Started by adding nucleotides at 37°C	M1 Theme: the reactions Instrument (change): with 2% perchloroacetic acid Predicate: stopped Event_Type: Data collection

Event 13	Event 14	Event 15	Event 16
M2 Theme: Nucleotides Instrument (change): by HPLC on an anion exchange column (a 10 × 0.46 mm SAX column from Rainin Instruments) Predicate: separated HPLC (high performance liquid chromatography) is a general method or instrument that used in experiments. an anion exchange column (a 10 × 0.46 mm SAX column from Rainin Instruments), this is the specific type of HPLC. We need to come up with a labelling for this kind of information ("based on the method of Hartwick and Brown [29]") Event_Type: Data collection	M2 Theme: Vanadyl and nucleotide solution We need to decide what label should we give this: according to Houseman et al. [21] Predicate: prepared Event_Type: Data collection	M1 Theme: Dissolved molecular oxygen Instrument(change): by purging with dry nitrogen gas Predicate: removed (prepared) Event_Type: Data collection	M1 Theme: the solvent – the buffer from the previous events containing the sCD39 The solute: sC39 Instrument (change): an YM30 centricon from Millipore Predicate: changed Changed to (we don't have a semantic role to label this kind of information): changed to 20 mM Hepes (pH8.0), 120 mM NaCl, 5 mM KCl Event_Type: Data collection

Preparation of VO^{2+} solution

12- Vanadyl and nucleotide solution were prepared according to Houseman et al. [21].
13- Dissolved molecular oxygen was removed from solutions by purging with dry nitrogen gas.
14- Stock vanadyl and nucleotide solution were thawed on ice, and mixed at 1:1 molar ratio by vigorous stirring.
15- Then VO^{2+}-nucleotide complexes were added to purified sCD39 at 1:1 molar ratio, mixed, and incubated for 5 minutes on ice before they were transferred into EPR tubes.
16- Once the samples were in EPR tubes, they were immediately frozen in liquid nitrogen, and stored in liquid nitrogen before using.

Event 17	Event 18	Event 19	Event 20
M1 Theme: Stock vanadyl and nucleotide solution Instrument (change): by vigorous stirring Predicate: mixed	M1 Theme: VO2+-nucleotide complexes Instrument (change): by vigorous stirring Predicate: added and mixed Event_Type: Data collection	M1 Theme: the samples Instrument (maintain): incubating on ice Predicate: incubated Time: for 5 minutes Event_Type: Data collection	M1 Theme: the samples Goal: into EPR tubes Predicate: transferred Event_Type: Data collection

EPR Measurement

1- CW-EPR experiments were carried out at X-band (9 GHz) using a Bruker 300E spectrometer with a TE102 rectangular standard cavity and a liquid nitrogen flow cryostat operating at 150 K.
2- Simulations of these EPR spectra were accomplished with the computer program QPOWA [30,31]).
3- To estimate the types of groups that serve as equatorial ligands to VO2+ in each condition, the observed values of $A_{||}$ derived from simulation of the EPR spectrum by QPOWA were compared with the coupling constants obtained from model studies [24,32] using:
   $A||calc = \Sigma\ niA||i/4$
   where i represents the different types of equatorial ligand donor groups, n_i (=1–4) is the number of ligands of type i, and $A_{||i}$ is the measured coupling constant for equatorial donor group i [24].
4- Similar equations were used to calculated $g_{||}$ from a given set of equatorial ligands for comparison with those derived experimentally.

Event 21	Event 22	Event 23	Event 24
M1 Theme: the samples Instrument (change): in liquid nitrogen Predicate: frozen Event_Type: Data collection	M1 Theme: the samples Instrument (maintain): in liquid nitrogen Predicate: stored Event_Type: Data collection	M1 Theme: CW-EPR experiments Instrument (measure): using a Bruker 300E spectrometer Predicate: carried out Protocol detail: with a TE102 rectangular standard cavity and a liquid nitrogen flow cryostat operating at 150 K Event_Type: Data collection	M1 Theme: Simulations of these EPR spectra Instrument (mathematical instrument): with the computer program QPOWA [30,31]) Predicate: accomplished Event_Type: Data analysis

Event 25	Event 26	Event 27
M1 Patient: the observed values of A|| derived from simulation of the EPR spectrum by QPOWA theme: with the coupling constants obtained from model studies [24,32] Predicate: compared Result: To estimate the types of groups that serve as equatorial ligands to VO2+ in each condition Event_Type: Data analysis	M1 Theme: g|| Instrument (mathematical instrument): Similar equations from a given set of equatorial ligands Predicate: calculated Event_Type: Data analysis	M1 Theme: g|| Predicate: compared Patient (theme): those derived experimentally Event_Type: Data analysis

Important notes in the paper:

Theme: Soluble CD39
Predicate: purified
Cause: as described [25] (with some modification) – for 2nd tier of information

Event_Type: Data collection

Additional information from the reference in [25]:
Purification of sCD39. 300 ml of conditioned medium was harvested 48 h after each transfer and, after removal of cells and debris, was concentrated to about 10 ml using an Amicon concentrator. The solution was applied to 12 ml of concanavalin A-Sepharose 4B resin (Sigma) equilibrated with Buffer A (25 mM Tris–HCl, pH 7.0, 80 g NaCl and 5 g KCl per liter) containing 1 mM CaCl2, 1 mM MgCl2, and 1 mM MnCl2. Buffer A containing 0.1 mM a-methylmannoside, 25 ml, was used to wash the column. Proteins were eluted with 30 ml of Buffer A including 1 M a-methylmannoside. An Amicon concentrator was used to replace Buffer A with 40 mM Tris–HCl (pH 7.5), 100 mM NaCl, and 0.5 mM CaCl2. The solution was loaded on a 4 ml Pro-bond nickel resin equilibrated with 24 mM KH2PO4, 16 mM K2HPO4, and 0.5 M NaCl (pH 7.8). The column was washed with 8 ml of 24 mM KH2PO4, 16 mM K2HPO4, 5 mM imidazole and 0.5 M NaCl (pH 6.0) to remove non-specifically bound proteins. sCD39 was eluted with the same buffer containing 500 mM imidazole. The elution buffer was changed to 40 mM Tris–HCl (pH 7.5), 100 mM NaCl, 0.5 mM CaCl2 with an Amicon concentrator. Cold saturated (NH4)2SO4 was added to a final concentration of 60% saturation. After overnight incubation at 4°C, the precipitated protein was removed by spinning for 5 min at 15,000g. The supernatant was adjusted to 80% saturated (NH4)2SO4 and incubated for several hours at 4°C to allow sCD39 to precipitate completely. The precipitated sCD39 was collected and re-suspended in about 500 ml of 40 mM Tris–HCl (pH 7.5), 0.5 mM CaCl2. This sample was loaded on a Superose-12HR gel filtration column from Pharmacia Biotech equilibrated with 40 mM Tris–HCl (pH7.5), 0.5 mM CaCl2. The major peak was collected, and concentrated to about 50 ml with a YM30 Centricon concentrator from Millipore.

“Collected” can be used as a vague term because there could be a different method of collecting and it does not tell us how it was collected.

An important aspect is to find the “action term” that the authors did or use to cause the event to happen for example “the collection of sCD39” or “the resuspension of sCD39”. How did they collect the sCD39? Or How to resuspend the sCD39?

One way of solving this aspect, is as Sandor suggesting which is to remove the “cause” and replace it with “condition” which can include the “instrument”. (SOLVED using the modified instrument as a semantic roles)

But the problem here is that the notion of “”instrument” in the linguistic domain could be different from the biomedical domain. So, we have to ensure that what we consider as an “instrument” from the linguistic sense should be similar to the biomedical one, so we can correctly identify the instruments in the texts.

Direct information: something like “in 20 mM HEPES-Tris (pH 7.0), 120 mM NaCl, and 5 mM KCl”

Indirect information: something like stating the temp, “at 37 c” which indicates the incubator

Note:

The annotator should go back to that reference and try to see the actual protocol which may include multiple instruments which necessary to understand what they are doing in this paper.

Or they can call it as reference type instrument.

1: Description of method; M2: Reference to a protocol; M3: Source of materials; M4: Background information

Processes:

1- The over-expression plasmid for L1, pUB5832, was digested with NdeI and HindIII, and the resulting ca. 900 bp piece was gel purified and ligated using T4 ligase into pUC19, which was also digested with NdeI and HindIII, to yield the cloning plasmid pL1PUC19.
2- Mutations were introduced into the L1 gene by using the overlap extension method of Ho et al.[60], as described previously [68].
3- The ca. 900 bp PCR products were digested with NdeI and HindIII and ligated into pUC19.
4- The DNA sequences were analyzed by the Biosynthesis and Sequencing Facility in the Department of Biological Chemistry at Johns Hopkins University.

Event 1	Event 2	Event 3	Event 4
M1 Patient: The over-expression plasmid for L1, pUB5832, Instrument (catalyst): with NdeI and HindIII Predicate: digested Event_Type: Data collection	M1 Patient: the resulting ca. 900 bp piece Predicate: gel purified Instrument (catalyst): (implied info.) Gel electrophoresis – usually is the cause of gel purification. (From the ontology which was not stated in the text!!!!) Event_Type: Data collection	M1 Patient: pUC19 Instrument (catalyst): with NdeI and HindIII Predicate: digested Event_Type: Data collection	M1 Patient: the resulting ca. 900 bp piece Instrument (catalyst): using T4 ligase goal: into pUC19 Predicate: ligated Result: to yield the cloning plasmid pL1PUC19 Comments: 1- using the same restricted enzymes to create the same complimentary ends. Event_Type: Data collection

Event 5	Event 6	Event 7	Event 8
M1 Theme: the mutated pL1PUC19 plasmid Instrument (catalyst): with NdeI and HindIII Predicate: digested Event_Type: Data collection	M1 Theme: the 900 bp, mutated L1 gene Predicate: gel purified Instrument (change): Gel electrophoresis Event_Type: Data collection	M1 Theme: the 900 bp, mutated L1 gene goal: into pET26b Predicate: ligated Instrument (catalyst): using T4 ligase (implied info.) Result: to create the mutant overexpression plasmids Event_Type: Data collection	M1 Theme: E. coli BL21(DE3)pLysS Instrument(change) : with the mutated over-expression plasmids Predicate: transformed (and small scale growth cultures were used) - this information is part of event 12 which gives interpretation of E.coli. Event_Type: Data collection

5- After confirmation of the sequence, the mutated pL1PUC19 plasmid was digested with NdeI and HindIII, and the 900 bp, mutated L1 gene was gel purified and ligated into pET26b to create the mutant overexpression plasmids.
6- To test for overexpression of the mutant enzymes, E. coli BL21(DE3)pLysS cells were transformed with the mutated over-expression plasmids, and small scale growth cultures were used [68].
7- Large-scale (4 L) preparations of the L1 mutants were performed as described previously [36].
8- Protein purity was ascertained by SDS-PAGE.
9- The concentrations of L1 and the mutants were determined by measuring the proteins' absorbance at 280 nm and using the published extinction coefficient of ε280 nm= 54,804 M-1•cm-1[36] or by using the method of Pace[69]

Event 9	Event 10	Event 11	Event 12
M2 Theme: Large-scale (4 L) preparations of the L1 mutants Predicate: performed Instrument(reference): as described previously [36]. Event_Type: Data collection	M1 Theme: Protein purity Instrument(change): by SDS-PAGE Predicate: ascertained Event_Type: Data collection	M1 Theme: The concentrations of L1 and the mutants Instrument (measure): by measuring the proteins' absorbance at 280 nm and using the published extinction coefficient of ε280 nm= 54,804 M-1•cm-1 [36] (spectrophotometry-implied info) Instrument (reference): (and) (or) by using the method of Pace[69] !! Predicate: determined Event_Type: Data collection Comment: We are not completely satisfied with the interpretation of this sentence because it is ambiguous However, we add the coordinator conjunction “and” to replace the “or” which make the confusion.	M1 Theme: the protein samples Instrument (change): buffer and semi-permeable membrane Predicate: dialyzed Protocol detail: Duration (time): over => (lasting) 96 hours Temp: at 4°C Buffer: 3 × 1 L of metal-free, 50 mM HEPES, pH 7.5 Event_Type: Data collection

10- Before metal analyses, the protein samples were dialyzed versus 3 × 1 L of metal-free, 50 mM HEPES, pH 7.5 over 96 hours at 4°C.
11- A Varian Inductively Coupled Plasma Spectrometer with atomic emission spectroscopy detection (ICP-AES) was used to determine metal content of multiple preparations of wild type L1 and L1 mutants.
12- Calibration curves were based on three standards and had correlation coefficient limits of at least 0.9950. (IMPLICIT INFORMATION REQUIRED ONTOLOGY)
13- The final dialysis buffer was used as a blank, and the Zn(II) content in the final dialysis buffers was shown to be < 0.5 μM (detection limit of ICP) in separate ICP measurements.
14- The emission line of 213.856 nm is the most intense for zinc and was used to determine the Zn content in the samples.
15- The errors in metal content data reflect the standard deviation (σn-1) of multiple enzyme preparations. (This is not an event, this is a result of an event) – Additional information
16- Steady-state kinetic assays were conducted at 25°C in 50 mM cacodylate buffer, pH 7.0, containing 100 μM $ZnCl_2$ on a HP 5480A diode array UV-Vis spectrophotometer at 25°C.

Event 13	Event 14	Event 15	Event 16
M1 Theme : metal content of multiple preparations of wild type L1 and L1 mutants Instrument (measure) : A Varian Inductively Coupled Plasma Spectrometer with atomic emission spectroscopy detection (ICP-AES) Predicate: used (determined) Event_Type: Data collection	M1 Theme: Curves Predicate: calibration (nominalized verb) Result: at least 0.9950 Instrument (measure): Spectrometer (This is an implied information not found in the text) Event_Type: Data collection	M1 Theme: Curves Predicate: calibration (nominalized verb) Result: at least 0.9950 Instrument (measure): Spectrometer (This is an implied information not found in the text) Event_Type: Data collection	M1 Theme: The final dialysis buffer Predicate: used Cause: The researchers Event_Type: Data analysis

Event 17	Event 18	Event 19	Event 20
M1 Theme: the Zn(II) content in the final dialysis buffers Instrument (measure): ICP (Spectrometer) Predicate: shown Result: to be < 0.5 µM (detection limit of ICP) in separate ICP measurements Event_Type: Data collection	M1 Theme: The emission line of 213.856 nm Predicate: used Result: to determine the Zn content in the samples Event_Type: Data analysis	M1 Theme: Steady-state kinetic assays Predicate: conducted (measured) Instrument (measure): on a HP 5480A diode array UV-Vis spectrophotometer at 25°C Protocol detail: Temp: 25°C Buffer: 50 mM cacodylate, pH 7.0 Cofactor: 100 µM ZnCl2 Instrument: on a HP 5480A diode array UV-Vis spectrophotometer Event_Type: Data collection Comments: all these parts of the cause and we considered them as the experimental settings.	M1 Theme: substrate concentrations Predicate: varied Cause: the researchers Result: 0.1 to 10 times the Km value Event_Type: data collection Even if the authors are the cause because the event involving manipulation of physical characteristics of data analysis.

17- When possible (stylistic comment to describe the connection between the following texts), substrate concentrations were varied between 0.1 to 10 times the Km value.
18- In kinetic studies using substrates with low Km values (cefoxitin, nitrocefin, and cephalothin) or with small Δε values (penicillin and ampicillin), we typically used substrate concentrations varied between ~ Km and 10 × Km and used as much of the ΔA versus time data (that was linear) as possible to determine the velocity.
19- Steady-state kinetics constants, Km and kcat, were determined by fitting initial velocity versus substrate concentration data directly to the Michaelis equation using CurveFit [36].
20- The reported errors reflect fitting uncertainties. (This is not an event)
21- All steady-state kinetic studies were performed in triplicate with recombinant L1 from at least three different enzyme preparations. (This is not an event) It is added detail.
22- Circular dichroism samples were prepared by dialyzing the purified enzyme samples versus 3 × 2 L of 5 mM phosphate buffer, pH 7.0 over six hours.
23- The samples were diluted with final dialysis buffer to ~75 µg/mL.
24- A JASCO J-810 CD spectropolarimeter operating at 25°C was used to collect CD spectra.
25- Rapid-scanning Vis spectra of nitrocefin hydrolysis by L1 and the L1 mutants were collected on a Applied Photophysics SX.18MV stopped-flow spectrophotometer equipped with an Applied Photophysics PD.l photodiode array detector and a 1 cm pathlength optical cell
26- A typical experiment consisted of 25 µM enzyme and 5 µM nitrocefin in 50 mM cacodylate buffer, pH 7.0 containing 100 µM ZnCl2, the reaction temperature was thermostated at 25°C, and the spectra were collected between 300 and 725 nm.
27- Data from at least three experiments were collected and averaged.
28- Absorbance data were converted to concentration data as described previously by McMannus and Crowder [39].
29- Stopped-flow fluorescence studies of nitrocefin hydrolysis by L1 were performed on an Applied Photophysics SX.18MV spectrophotometer, using an excitation wavelength of 295 nm and a WG320 nm cut-off filter on the photomultiplier.
30- These experiments were conducted at 10°C using the same buffer in the rapidscanning Vis studies.
31- Fluorescence data were fitted to kobs = {(kf [S]) / KS + [S])} + kr as described previously [40] or to kobs = kf [S] + kr by using CurveFit v. 1.0.

Event 21	Event 22	Event 23	Event 24
M1 Agent: We (the authors) Predicate: determine Theme: the velocity Instrument (change): substrate concentrations varied between ~ Km and 10 × Km and used as much of the ΔA versus time data (that was linear) as possible Comments: Km is the substrate concertation when the enzyme is doing in half Vmax (which is rate of reaction). A is the absorbance which measure how much light have been absorbed in molecular. Event_Type: data collection	M2 Theme: Steady-state kinetics constants, Km and kcat Instrument (mathematical): by fitting initial velocity versus substrate concentration data directly to the Michaelis equation using CurveFit [36] Predicate: determined Event_Type: Data analysis	M1 Theme: Circular dichroism samples Instrument(change): by dialyzing the purified enzyme samples versus 3 × 2 L of 5 mM phosphate buffer, pH 7.0 over six hours Predicate: dialyzed Protocol detail: Buffer:3 × 2 L of 5 mM phosphate buffer, pH 7.0 Time: six hours Event_Type: Data collection	M1 Theme: The samples Instrument(change): with final dialysis buffer to ~75 µg/mL Predicate: diluted Event_Type: Data collection

Event 25	Event 26	Event 27	Event 28
M1 Theme: CD spectra Predicate: collect (measure) Instrument(measure): A JASCO J-810 CD spectropolarimeter operating at 25°C 3 × 2 L of 5 mM phosphate buffer, pH 7.0 Temp: at 25°C Buffer: 3 × 2 L of 5 mM phosphate buffer, pH 7.0 (the buffer is implied by the event 30 because it stated the samples were diluted) Event_Type: Data collection	M1 Theme: Rapid-scanning Vis spectra Instrument(measure): a Applied Photophysics SX.18MV stopped-flow spectrophotometer equipped with an Applied Photophysics PD.l photodiode array detector and a 1 cm pathlength optical cell Predicate: collected Event_Type: Data collection M1 Nested event (bio-event) of event 32 Theme: nitrocefin Predicate: hydrolysis (nominalized) Cause: by L1 and the L1 mutants Event_Type: Data collection	M1 Theme: typical experiment Predicate: consisted (cause)of25 µM enzyme and 5 µM nitrocefin in 50 mM cacodylate buffer, pH 7.0 containing 100 µM ZnCl2 Not an event (Extra information for the previous event) It could be protocol detail.	M1 Theme: the reaction temperature Predicate: thermostated At 25°C, Not an event (Extra information for the previous event) It could be protocol detail.

Event 29	Event 30	Event 31	Event 33
M1 Theme: and the spectra Predicate: collected between 300 and 725 Nm Not an event (Extra information for the previous event) It could be protocol detail.	M1 Theme: Data from at least three experiments Predicate: averaged Cause: the researchers Event_Type: Data analysis	M1 Theme: Absorbance data Cause: the researchers Result: to concentration data Predicate: converted Event_Type: Data analysis **Event 32** M1 Theme: Fluorescence data Instrument(mathematical): to kobs = {(kf [S]) / KS + [S])} + kr as described previously [40] OR to kobs = kf [S] + kr by using CurveFit v. 1.0. Predicate: Fitted Event_Type: Data analysis	M1 Theme: Stopped-flow fluorescence studies of nitrocefin hydrolysis by L1 Predicate: performed (measured) Instrument (measure): on an Applied Photophysics SX.18MV spectrophotometer, using an excitation wavelength of 295 nm and a WG320 nm cut-off filter on the photomultiplier. Event_Type: Data collection These experiments were conducted at 10°C using the same buffer in the rapidscanning Vis studies. (Extra information)

Important comments:

- “By” could be used to indicate the use of instrument.
- The verb “used” almost is used as coordinator verb and the main verb should be following the verb in infinite form “to wash” or as nominalized verb such as “activation”
- Are all nested events should be “bio-events”?
- Important note too is to separate the events into two different categories: 1- data collection 2- data analysis.
- NdeI and HindIII used in both digestions because they will give us the same complementary sticky ends
- The over expression plasmid means it has the machinery (actively promoter and enhancer) that can express the gene in great quantity
- Most of the time in data analysis event the researchers are the cause.
- If removing a gene from a plasmid and putting that gene into different plasmid, so the digestion process has to happen before the ligation

Appendix C

XML Schema for Semantic Roles and Rhetorical Moves

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!— XSchema deffinition for Rhetorical Moves —>
  <element name="ArgMoves">
    <complexType>
      <attribute name="Type" use="required" value="other" >
        <simpleType>
          <restriction base="string">
            <enumeration value="Description_method"/>
            <enumeration value="Reference_to_method"/>
            <enumeration value="Source_materials"/>
            <enumeration value="Background_information"/>
          </restriction>
        </simpleType>
      </attribute>
      <attribute name="Relation" use="required" value="other" >
        <simpleType>
          <restriction base="integer">
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

Figure C.1: XML Schema for Rhetorical Moves

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!— XSchema deffinition for Main Semantic Roles—>
  <element name="SemanticRole">
    <complexType>
      <attribute name="Type" use="required" value="other" >
        <simpleType>
          <restriction base="string">
            <enumeration value="Theme"/>
            <enumeration value="Agent"/>
            <enumeration value="Patient"/>
       <enumeration value="Goal:Physical"/>
       <enumeration value="Goal:Purpose"/>
            <enumeration value="Location"/>
            <enumeration value="Factitive"/>
          </restriction>
        </simpleType>
      </attribute>
      <attribute name="Relation" use="required" value="other" >
        <simpleType>
          <restriction base="integer">
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

Figure C.2: XML Schema for Semantic Roles

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!— XSchema deffinition for Instrument Types —>
  <element name="SemanticRole:Instrument">
    <complexType>
      <attribute name="Type" use="required" value="other" >
        <simpleType>
          <restriction base="string">
            <enumeration value="Instrument:Change"/>
            <enumeration value="Instrument:Measure"/>
            <enumeration value="Instrument:Observe"/>
            <enumeration value="Instrument:Maintain"/>
            <enumeration value="Instrument:Catalyst"/>
            <enumeration value="Instrument:Mathematical"/>
            <enumeration value="Instrument:Reference"/>
          </restriction>
        </simpleType>
      </attribute>
      <attribute name="Relation" use="required" value="other" >
        <simpleType>
          <restriction base="integer">
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

Figure C.3: XML Schema for Semantic Role (Instrument)

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!— XSchema deffinition for Protocol Detail Information —>
  <element name="SemanticRole:Protocol_Detail">
    <complexType>
      <attribute name="Type" use="required" value="other" >
        <simpleType>
          <restriction base="string">
            <enumeration value="ProtocolDetail:Temp"/>
            <enumeration value="ProtocolDetail:Time"/>
            <enumeration value="ProtocolDetail:Repetition"/>
            <enumeration value="ProtocolDetail:Condition"/>
            <enumeration value="ProtocolDetail:Cofactor"/>
            <enumeration value="ProtocolDetail:Buffer"/>
          </restriction>
        </simpleType>
      </attribute>
      <attribute name="Relation" use="required" value="other" >
        <simpleType>
          <restriction base="integer">
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

Figure C.4: XML Schema for Semantic Role (Protocol Detail)

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!-- XSchema deffinition for verbs -->
  <element name="Predicate">
    <complexType>
      <attribute name="Type" use="required" value="other" >
        <simpleType>
          <restriction base="string">
            <enumeration value="Verb"/>
            <enumeration value="Nominlised_Verb"/>
          </restriction>
        </simpleType>
      </attribute>
    </complexType>
  </element>
</schema>
```

Figure C.5: XML Schema for Predicates

Appendix D

Steps of Alkaline Agarose Gel Electrophoresis

Alkaline Agarose Gel Electrophoresis

1. Prepare the agarose solution

1.1 Adding the appropriate amount of powdered agarose to a measured quantity of H2O in either:

1.1.1 An Erlenmeyer flask (Container 1)

1.1.1.1 Loosely plug the neck of the Erlenmeyer flask with Kimwipes

1.1.2 OR a glass bottle (Container 1)

1.1.1.2 Make sure that the cap is loose

1.2 Heat the slurry (Item1) in (Conatiner1) for the minimum time required to allow all of the grains of agarose to dissolve using either:

1.2.1 A boiling-water bath

1.1.1.3 Check that the volume of the solution (Item 1) has not been decreased by evaporation during boiling in (Container 1):

1.1.1.3.1 if yes: replenish with H2O in (Container 1)

1.1.1.3.2 If no: do not add H2O in (Container 1)

1.2.2 OR a microwave oven

1.2.2.1 Check that the volume of the solution (Item 1) has not been decreased by evaporation during boiling in (Container 1):

1.2.2.1.1 if yes: replenish with H2O in (Container 1)

1.2.2.1.2 If no: do not add H2O in (Container 1)

1.3 Cool the clear solution (Item 1) to 55 C.

1.3.1 Add 0.1 volume of 10x alkaline agarose gel electrophoresis buffer in (Container 1)

1.3.2 And immediately pour the gel (Item 1) into mold (Container 2)

1.4 After the gel (Item 1) is completely set

1.4.1 Mount it (Item 1) in the electrophoresis tank (Container 3)

1.4.2 Add freshly made 1x alkaline electrophoresis buffer until the gel (Item 1) is just covered.

2. Prepare DNA samples

2.1 Collect the DNA samples (Item 2) by standard precipitation with ethanol

2.2 Dissolve the damp precipitates of DNA (Item 2) in 10-20 µl of 1x gel buffer. (Item 3)

2.3 Add 0.2 volume of 6x alkaline gel-loading buffer.

2.3.1 It is important to chelate all Mg2+ with EDTA before adjusting the electrophoresis samples to alkaline conditions.

3. Initiate the electrophoresis

3.1 Load the DNA samples dissolved in 6x alkaline gel-loading buffer into the wells of the gel (container 3)

3.2 Start the electrophoresis at <3.5 V/cm when the bromocresol green has migrated into the gel approx. 0.5-1 cm; Turn off the power supply, and place a glass plate on top of the gel in (Container 3) and then continue electrophoresis until the bromocresol green has migrated approximately two thirds of the length of the gel in (container 3).

4. Finalize the experiment

4.1 Process the gel according to one of the procedures either Southern hybridization by:

4.1.1 Transfer the DNA either:

4.1.1.1 Directly (without soaking the gel) from the alkaline agarose gel to a charged nylon membrane. Please see Southern Blotting: Capillary Transfer of DNA to Membranes

4.1.1.2 OR after soaking the gel in neutralizing solution for 45 minutes at room temperature to either:

4.1.1.2.1 An uncharged nitrocellulose as described in Southern Blotting: Capillary Transfer of DNA to Membranes

4.1.1.2.2 OR nylon membrane as described in Southern Blotting: Capillary Transfer of DNA to Membranes

4.1.2 Detect the target sequences in the immobilized DNA by hybridization to an appropriate labeled probe. Please see Southern Hybridization of Radiolabeled Probes to Nucleic Acids Immobilized on Membranes

4.2 OR Staining

4.2.1 Soak the gel in neutralizing solution for 45 minutes at room temperature.

4.2.1.1 Stain the neutralized gel with 0.5 µg/ml ethidium bromide in 1x TAE or with SYBR Gold.

4.2.1.1.1 A band of interest can be sliced from the gel and subsequently eluted by one of the procedures described Recovery of DNA from Agarose Gels

Appendix E

List of Procedural Verbs

Verb	Definition and their usage
Agitated	**Definition:** To stir. Other synonyms might be: mix, break-up/apart, invert, shake, swish, pipette up and down. Can be part of almost any process or procedure: you do it making a gel, preparing a solution, miniprep, transferring cells from one plate to another, etc. **Why is it done:** In any case where you want uniform distribution of materials in a liquid. You might do it to it to: Solutions, gels, suspensions, mixtures, liquid cultures, buffers. **How is it done:** Revolutions per minute (rpm), stir rod, stir bar (could be on low, medium, high intensity), shaken, inverted (inverting a microfuge tube is often done to stir with as low intensity as possible), pipette up and down, swishing a plate, gently, mildly, vigorously, intensely, at a particular temperature.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Amplified	**Definition:** Amplification of DNA means reproducing a piece of DNA (and its sequence) many times to yield exponentially more copies of the DNA sequence. Synonyms: copied, reproduced, replicated. **Why is it done:** It can be done for two main reasons. 1) To have lots of sample to work with (for whatever is being done with it). 2) To have enough to be visualized on gel (visualized under UV light it can be seen but only if there is enough). You might do it to: cDNA, a plasmid, an oligonucleotide, a gene, a sequence, DNA library, a site or a region of a gene, genome, DNA fragment. **How is it done:** Polymerase Chain Reaction PCR, a particular number of 'cycles' (referring to PCR), in a cloning vector, in a thermal cycler (PCR machine)(may be different brands), at a particular set of temperatures for particular sets of time depending on, from specific DNA or RNA primers, RT-PCR .
Annealed	**Definition:** Single strands of DNA (usually they have to be complementary or close to it) electrostatically coming together to form double or paired strands. **Why is it done:** It happens during the PCR process where small single stranded DNA primers anneal to to the two original DNA strands separated by heat. In this process it comes after denaturation, and before elongation/extension. But annealing can be done whenever there are single complementary DNA strands. For example sometimes when ordering oligonucleotide stocks they come in single strands and you have to anneal them before use. You might do it to: oligonucleotides (oligos), DNA, RNA, ssDNA/ssRNA (single stranded), complements/complementary strands, strands, denatured DNA. **How is it done:** At a particular temperature and for a particular amount of time, (A lot of the same key words as for verb "amplified" since both are part of PCR), in a buffer.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Bind	**Definition:** this is more of a chemistry term that means any sort of interaction between atoms or molecules that involves them coming together and sticking to one another. There are many different kinds of bonds/binding and some are stronger and some are weaker. The strongest bonds are ionic bonds, but those are almost never of interest in biochemistry or molecular biology. Covalent bonds are slightly weaker than ionic bonds and are what hold most organic molecules together. For example, in a glucose molecule there are a number of carbon, oxygen and hydrogen atoms. Those are all non-metals meaning they covalent bond to one another. There are weaker electrostatic interactions than that such as hydrogen bonds which are very important for proteinligand binding as often it is hydrogen bonds that form the active or binding site. Binding can be done actively by researchers for different reasons but it also occurs non-stop in nature, so at times, binding is the subject of study rather than the protocol (like in signaling pathways). Synonyms: bond, bonded. **Why is it done:** There are many reasons one would want to molecules to bind. You could put a molecular marker on a protein to track it, you might want to see a potential substrate drug bind to the protein of interest to inhibit it or adjust it slightly to try to optimize the binding, or you might be using it to purify a protein in a column. Binding is literally everywhere so it is difficult to identify all the protocols it might be a part of. You might do it to: proteins, ligands, enzymes, substrates, DNA, RNA, molecules, cofactors, beads, membranes. **How it is done:** By forming immunocomplexes, electrostatic interactions.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Biotinylated	**Definition:** The covalent bonding of a biotin (which is a molecule) to something (some other biomolecule, usually a protein). **Why is it done:** 1) as a marker for detection (as mentioned above in the binding section), 2) as a tag for protein purification (affinity chromatography also mentioned above). Biotin has very high affinity for streptavidin and avidin, this means that you can have biotin bind a protein, release that protein into the cell or a buffer, then you can always locate/isolate that protein again using avidin to bind the biotin that is bound to the protein of interest. You might do it to: any biomolecule, usually a protein. **How it is done:** In a buffer at a particular temperature, chemically, enzymatically, primary amine biotinylation, sulfhydryl biotinylation, carboxyl biotinylation, glycoprotein biotinylation, oligonucleotide biotinylation, non-specific biotinylation. As with binding, the specific ways that biotin binds to something are vast and is more in the realm of chemistry than biochemistry. The two general methods it is a part of would be good to analyze: (affinity chromatography and biotin protein detection).
Carboxymethylated	**Definition:** The addition of a carboxymethyl group (a molecule) into a compound. **Why is it done:** Commonly done as a step during protein sequencing when a protein has disulphide bonds (covalent bonds between two cysteines (an amino acid residue) of a single protein). First there is a 'reduction' step which cleaves the cysteine bond, however the residues remain highly reactive which can make the sequencing difficult. Carboxymethylation of a reduced cystein residue bind to it and acts as a cap to prevent it from reacting. You might do it to: a protein or peptide. **How it is done:** In a particular buffer at a particular temperature (can vary greatly), done after reduction and before protein sequencing.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Centrifuged	**Definition:** A machine that rotates samples rapidly to allow the contents of a solution/suspension to sediment. Supposed to achieve what gravity would over time in a short period of time. The verb centrifuged just meant to put the subject through a centrifugation (using the machine). It is a part of many protocols for many reasons. There are different types of centrifugation: high-speed, microcentrifugation, ultracentrifugation, density gradient centrifugation, differential centrifugation, but all of these involve the use of a centrifuge that spins and all of them are done with the ultimate goal of seperating species. Synonyms: Spun, pelleted (although there may be different ways of doing this), sedimented. **Why is it done:** To separate different species (things) in the same medium (solution/suspension) by making one a solid pellet and then discarding either the pellet or the supernatent fluid. Can be done to wash a material of interest in a spin column with buffer. You might do it to: solutions, suspensions, cells, buffers, a supernatent, a fraction, cell lysate/lysates, soluble material/ material, medium. Basically any word for the subject of interest in a liquid or they just refer to the subject of interest without mentioning the liquid it is in contact with but it is implied. **How is it done:** At a certain RPM, at a certain number of "g's" in a certain buffer at a certain temperature, for a certain amount of time, in a particular instrument, a certain number of times, using a density gradient column, using a spin column, using a microfuge tube.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Cloned	**Definition:** To make another copy of DNA/RNA. Cloning into has a different meaning however and is associated with the term molecular cloning which is where DNA is inserted into a replicating vehicle like a viral vector or a plasmid. So you can clone something INTO a cloning vector and then have it cloned to make more copies. Synonyms for cloning: amplified, replicated. **Why is it done:** To acquire more copies of a segment of DNA, to have DNA in a recombinant vector for whatever reason: sequencing, increased expression, to prepare plasmid for insertion into something else (organism), to add something to the N-terminal or c-terminal end of an existing gene, to create fusion proteins. You might do it to: cDNA, DNA fragment, gene, RNA, coding region, spacer region, a plasmid, an oligonucleotide, a sequence, DNA library, a site or a region of a gene, promoter, terminator, a domain. **How is it done:** Using/at restriction sites (particular sites), ligation/ligases, in frame, in a vector, in a particular buffer, at a particular temperature, for a particular time, as fusion proteins, between two genetic elements, restriction enzyme techniques.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Collected	**Definition:** This has no specific biological meaning. It just means to gather or obtain a thing. That is how it is used in the sentences you provided although it will often refer to the fancy biochemical method for how something was collected, but each of those things will have a term of its own if you look at the sentences. Synonyms are "harvested" and "extracted". **Why is it done:** To gather or obtain a thing, often after or before or after it has undergone some procedure. You might do it to: literally anything: proteins, DNA, RNA, molecules, atoms. **How is it done:** Look at the full sentence for the relevant context. Some sentences refer to a BioRad (Model 2110) fraction collector which is machine for collection after chromatography while other use a filter, and others use centrifugation. But again, there are limitless conceivable ways that something could be collected, I think that the important thing is just to recognize that it is just the English word collected and that the subject and the process are referred to usually in the same sentence.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Concentrated	**Definition:** A chemistry term referring to how much of something there is per volume (similar to density which is defined more specifically as mass per unit volume). Molar (M) is a unit of concentration, meaning moles per litre. **Why is it done:** It is important for almost everything, pH is a scale that measures the concentration of hydrogen atoms in a solution, all buffers will have one or more concentrations listed as it is at those particular abundance of atoms/molecules in an experiment that the process needs to be carried out. Many acids and bases come in extremely high concentrations so that you get more for your money by just diluting and using what you need. Concentration of a competitive inhibitor is what determines if it will out-compete the substrate for protein binding. The higher concentration will spend more time bound to the protein (because of probability). And concentration also determines things like equilibrium across a permeable membrane and is what allows the ion channels in neurons to send brain signals. The topic is absolutely everywhere and absolutely important. It is not really a protocol however (other than changing concentration). Just a thing that exists to be aware of. It might apply to: a solution, a suspension, a buffer, some medium. **How is it done:** Changing the concentration of something is done using this equation C1V1=C2V2 where is original concentration times the original volume must be equal to the final concentration times the final volume. So for example if I know that I want a 10 mL of a 1M solution of glucose and the stock is 10M then the variable is the original volume. So it rearranges to)(C2*V2)/C1 = V1 and I can solve for the volume of the 10 M solution to add to water to get 10 mL. Or just rearrange to find whatever I don't know based on what I need to know.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Conjugated	**Definition:** The direct transfer of DNA (plasmid that is separate from the main bacterial chromosome) between bacterial cells. It is done as a form of sexual reproduction between bacteria. Although they can reproduce asexually, transferring DNA can give them some variation and prepare them for things that other bacteria have experienced. However, all your sentences that I can see are talking about conjugated antibodies, which are antibodies linked to a molecular label that are used for detection in many assay techniques. An example of a molecular label might be something with fluorescence that you can use to detect where your antibodies have bound and therefor where your protein of interest is. **Why is it done:** To detect the location of a protein of interest within a cell using fluorescence microscopy. Or more generally, to detect something that the antibody binds to, using various detection assays. (Similar to one of the uses of biotin). You might do it to: antibodies, or bacteria under the other meaning. **How is it done:** To a protein or biomolecule, for a certain period of time, at a certain temperature in a certain buffer, often done in conjunction with horseradish peroxidase (which causes luminescence), usually listed as [LABEL] conjugated, anti-[PROTEIN OF INTEREST] to describe what the conjugated antibody actually is. Antibody can be conjugated to other materials.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Cultured	**Definition:** It means grown or maintain cells in specific conditions. Synonyms: plated, grown, incubated. **Why is it done:** To have cells to work with or to allow for some process of interest inside the cells to take place. You might do it to: bacteria, tissue cells, any type of cells, fungi. **How is it done:** In a specific media, at a specific temperature, with specific nutrients, for a certain amount of time, at a certain humidity, in a specific atmosphere. On 6 well plates, 10 cm plates, 24 well plates, 48 well plates. Done before and after transfection/transformation. There are many different conditions for cultures as there are different conditions for each cell type, what organ its from/organism its from, in addition to what process you are planning to do to the cells. For example, some tough algae should be grown in stressful conditions to the cells (e.g., High salt concentration) as that makes them more receptive to transformation. So not always are culture conditions optimal for growth, some conditions are optimal for other things and conditions can be changed accordingly.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Denatured	**Definition:** It means that something is being altered from its natural or functional state. In proteins or RNA this generally means destroying their function by heating them. But in DNA it generally refers to the separation of double strands. **Why is it done:** In many cases in proteins in RNA denaturing is something to be avoided but it is also used in protein sequencing or in SDS gels where a proteins size is to be determined. Denaturing a protein linearizes it making its run on a gel separate it by mass and prevent 3D effects from altering the distance it travels. In DNA denaturing is done as a step during PCR when the temperature is 95 degrees. This causes the two strands to split apart, the step comes before annealing. You might do it to: Protein, DNA, RNA. **How is it done:** At a particular high temperature, in a certain buffer, by SDS polyacrylamide in a gel, for a certain amount of time, using urea, might be done by boiling, with other various denaturing agents.
Digested	**Definition:** In molecular biology this means cut or cleaved (usually DNA). But it can also mean in certain contexts that a large end of DNA was destroyed or removed or that something was destroyed entirely. **Why is it done:** To ligate something into a plasmid, subcloning, molecular cloning, preparing a sample for something. You might do it to: DNA, RNA, Protein. (In order of what you will see most frequently). **How is it done:** with a specific restriction enzyme, at a particular cut site, from a certain 'end', at a particular temperature, for a certain amount of time, in a particular buffer, at a particular sequence/coding or non-coding region. Often digested will be followed by "with" which describes the restriction enzyme or enzymes responsible for performing the digestion.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Diluted	**Definition:** This means to lower the concentration of something. **Why is it done:** In any case where one might want a smaller concentration of something. You might do it to: solutions or suspensions containing a particular molecule, DNA, RNA, protein, atoms, antibodies. **How is it done:** Generally by taking a small volume of a solution and pipetting it into a larger volume of plain water, some other liquid, or a particular buffer. From a certain concentration to a certain concentration. Diluted a certain "fold" which refers to the order of magnitude. So if one dilute something 3 fold that means she or he performed 3 "serial dilutions" where she or he took 1 unit of the original liquid and added it to 10 units of water/buffer, then took 1 unit from that newly diluted solution and put that into 10 units of fresh water/buffer. "Fold" refers to the number of times one repeats that serial dilution process. Some times expressed as diluted 1:1, or 1:X as a ratio of the number of parts of the subject for 10 parts of the liquid/or whatever they are comparing it too.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Eluted	**Definition:** To remove something (protein, molecule, DNA) from a container by washing it with a solvent. **Why is it done:** Done in chromatography frequently to remove the purified protein from a column as well as during miniprep to remove DNA after. (So as one may see it often follows some purification step where washes were involved because the protein or DNA is being held in place by something (beads it is bound to for proteins and a filter for DNA miniprep) and then during the elution step some chemical or buffer and enzyme change finally allows it to leave its containment). You could also describe the junk that you don't want that gets thrown out in the washes as having been eluted however, so long as liquid/solution that is passing through is causing the thing to leave containment in the column/tube or whatever it may be. You might do it to: protein, DNA, RNA, molecule, atom, antibodies, anything. **How is it done:** With a particular buffer/liquid, eluting a particular volume or with a particular volumeoften determines the concentration of the subject which can be important for the steps to come. For example. After a miniprep of DNA that someone plans to transfect she or he wants it to be in as high concentration like 500 nM, so she or he elutes a DNA sample in 40 uL from a spin column instead of 50 uL. After HPLC (High Performance Liquid Chromatography which is a kind of chromatography). Usually elution occurs for a certain amount of time, at a certain temperature (although this is less important for elution so usually room temperature unless otherwise specified). Using a salt gradient (which makes thing elute in order of hydrophobicity). Different conditions such as buffers or filters can change what elutes first and elutions are collected in separate containers based on the time when they were eluted (because under a particular set of conditions different things elute faster/slower.) So is sometimes used as a crude method of separating many things.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Equilibrated	**Definition:** To bring something to or maintain an equilibrium – which in chemistry refers to species remaining in roughly the same concentrations despite changing state or transforming to different species. Almost every molecule/atom in the environment exists in a state of equilibrium, even water splits into oxygen and hydrogen naturally at a very low rate because without a catalyst that dissociation is very unlikely while the association reaction has a much higher rate. But that rate is constant under a certain temperature and pressure, changing conditions can change an equilibrium. **Why is it done:** Important for buffers and for a lot of procedures to have conditions that are constant. You might do it to: species in a buffer, solution, suspension, something in some medium. **How is it done:** It is something that happens on its own once a set of conditions is maintained, but it cannot happen when conditions are changing. This will therefore be done in a particular buffer at a particular temperature for a different amount of time. And in some cases at a specific humidity and atmosphere. In the general case in the sentences it appears to be a fancy way of letting you know that a new solution or buffer was allowed to sit for a minute before being further used.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Harvested	**Definition:** To collect cells or organisms or the medium they grow upon for further use (analysis or extraction). **Why is it done:** Generally after an experiment involving cells so that you can take a closer look at them or break them open and examine contents. You might do it to: cells, organisms, media. **How is it done:** Many ways, in a particular buffer, a lysis buffer containing PBS or just PBS itself, which cleaves the cells connection to the plate it whatever surface it is growing on, by centrifugation in which case at a particular RPM for a particular amount of time, done after incubating or experimenting for a certain amount of time. Often followed by resuspention in a buffer.
Impregnated	**Definition:** To put something inside something. **Why is it done:** When something needs to be inside a medium or some component of a mechanism in order to carry out a procedure. In the one sentence you provided it is being done to TLC plates as part of a PI3K activity assay. You might do it to: a gel, metallocene, a substrate, thin layer chromotography (TLC) plates. **How is it done:** Seems to be something that a company does to the equipment before hand a lot of the time. I found a document that lists methods for impregnating TLC plates and they include: spraying, dipping, pre-development, and mixed phases. But in the papers I've seen that use the word it usually describes some material they used that they purchased, not something they did themselves. For example: a [BLANK] impregnated [BLANK]. The first part of the sentence tells you what is being put in and the second part tells you what it is being put into. Unless you see: [BLANK] WAS impregnated WITH - [BLANK]. In which case the first blank is the medium that is having something put in it and the second blank is what is being put in.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Incubated	**Definition:** To maintain or grow something at a favorable temperature and other conditions that promote development. Generally refers to cultures of cells but could also be a tissue or organism. Occasionally referred to as "Overnight". **Why is it done:** To maintain or grow cells/organisms during a procedure or for procedures to come. Can also be done to carry out specific enzymatic reactions such as digestion (because enzymes may work best in a certain buffer and temperature. You might do it to: cells, tissue, organisms, a solution, a suspension, gels, a chamber, proteins, DNA, enzymes. **How is it done:** At a certain temperature for a certain amount of time, in a particular buffer or media, a certain volume, at a particular pH, with particular co-factors, in a solution, in a particular atmosphere (rarely), "overnight" indication a non-specific amount of time approximately 12 hours, in a medium.
Injected	**Definition:** To introduce something into something often after applying a force needed to enter and releasing the subject into it. **Why is it done:** To transport some subject of interest into an organism or structure that it would otherwise not be able to enter or diffuse into. Often done to introduce a disease or a drug to a mouse or rat as part of an experiment. You might do it to: organisms, gels, some specific organ or tissue of an organism, an organ or tissue growing without an organism, an oocyte/egg/embryo/cell, mouse, rat, animals, you might inject serum, a drug, a disease, a gene, anything. **How is it done:** With a certain volume of the subject, subcutaneously, a certain mass of the subject, into a column, into a HPLC system (you inject the sample you want to analyze with a needle), into a titration experiment, injected into a buffer via a pipette.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Ligated	**Definition:** DNA being stuck together at a break point(can be sticky ended or blunt ended). Note this is a break in the linear chain of nucleotides, not to be confused with annealing which is two unbroken linear strands joining together. Not always but can be sometimes synonymous with molecular cloning. **Why is it done:** To add something to something: ex. Add a gene to a plasmid or repair a breakage. You might do it to: DNA, RNA, plasmid. **How is it done:** Often done into plasmids (as molecular cloning) or into restriction endonuclease cut sites, can also be done onto the ends of linear DNA, between to restriction fragment cut sites, in a particular buffer at a particular temperature for a particular amount of time, using a particular ligase (enzyme that catalyses the ligation reaction), into a(n) (expression/cloning) vector, to or with an adaptor or cofactor, "upstream" or "downstream" from some genetic element such as: Promotor, terminator, enhancer, gene, intron, exon, and sequence.
Lysed	**Definition:** The destruction of a cell wall or cell membrane (Lysis). **Why is it done:** To extract something from the cell: ex. DNA. Or to kill cells that are unwanted (cancer). You might do it to: cells of any kind. The resulting solution is called a lysate and things of interest can be purified from it (proteins/DNA). **How is it done:** At a certain temperature, in a certain pH, in a certain buffer, over a certain amount of time, using a particular lysing reagent, in a "lysis buffer", on ice, at a particular volume, in a glass bead mill, in a French press (cells pop with physical force), via sonication (popping cells with sound waves), "vortexing" or just shaking them really fast can also lyse cells with physical force.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Phosphorylated	**Definition:** It means to attach a phosphate group to something. A phosphate group is a phosphorus bonded to four oxygen atoms. **Why is it done:** Occurs in many many signaling pathways so most often it is something being analyzed rather than the researchers actively phosphorylating something. Although in some cases it appears that beads or gels containing phosphorylated compounds are used for purification or proteolytic digestion, mass spectrometric analysis, and peptide sequencing. Another common example in nature is ATP (adenosine triphosphate) which is a molecule that allows for the transfer of energy around the cell vis the transfer of a phosphate group. You might do it to: any molecule, protein, and less often DNA. **How is it done:** A "kinase" is an enzyme that phosphorylates its substrate. The conditions of phosphorylation depends on the substrates and kinases but to have a kinase phosphorylate something in vitro it will be incubated in a specific buffer at a specific temperature for a specific amount of time. Phosphorylation (one category of binding reactions) can be measured a number of ways including: acid precipitation assay, mass spectroscopy, multi-analyse profiling, intracellular flow cytometry, enzyme-linked immunosorbent assay (ELISA), cell based ELISA, western blot, kinase activity assay, phospho-specific antibody development.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Pooled	**Definition:** Means combined or accumulated. **Why is it done:** Could be used in any context where two things are being poured into one another or when data from different experiments is being compiled for greater understanding. You might do it to: mixtures or solutions, results, or data from different experiments, fractions, buffers, proteins (in a solution but might not say that). **How is it done:** Context dependent, if referring to biochemical mixtures or solutions or materials it usually just means poured together. But it can mean just compiled in some way on a computer. Often followed by diluted or concentration (so some change in the concentration of the species of interest). There are no particular conditions under which it might occur.
Purified	**Definition:** To remove all of the things you do not want from a solution/suspension/medium while leaving only the subject of interest. **Why is it done:** In order to have a high purity of the subject of interest so that you know you are only looking at the thing you are interested in and nothing else, as other things might alter experiments. And secondly to obtain a high concentration of the subject of interest to increase the efficacy of an experiment that uses the subject. You might do it to: DNA, RNA, Proteins, Ligands, literally anything. If there is something in the cell you want to isolate or even a cell in a group of cells you purify it. **How is it done:** There are many different purification methods based on what is being purified: Centrifugation with ez-10 spin column or phenol chloroform extraction for DNA, or chromatography columns for proteins for example. So the conditions required can be vastly different.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Quenched	**Definition:** To stop or make inert. **Why is it done:** To silence the activity of a protein or enzyme for whatever reason. Often to stop a reaction at a particular amount of time to measure the proportion of reactants and products after a particular amount of time (could be used to measure rate of reaction). You might do it to: a protein or enzyme, a reaction. **How is it done:** With some chemical inhibitor. Structure: [BLANK] was quenched with [BLANK] where the first blank is the protein or enzyme and the second blank is the inhibitor, or quenching agent. In a "stop buffer".
Resuspended	**Definition:** To make solid particles or cells and introduce them to a suspension. **Why is it done:** For something like proteins or molecules or precipitates it is part of a purification process after precipitation out of a suspension. For example bacterial cells that were incubated overnight are precipitated by spinning quickly in a centrifuge then resuspended in lysis buffer. DNA can also be precipitated and resuspended in this manner for purification purposes. For cells generally as well, after harvesting you resuspend them in a buffer. With live cells the term usually suggests that they were originally suspended in nutrient rich media and are now being resuspended in a buffer where experiments will be performed on them. But it could technically be used to describe any resuspension from any medium to another. You might do it to: cells, molecules, particles, precipitates. As long as it is insoluble and was suspended previously you can suspend it. **How is it done:** In a certain buffer, sometimes then incubated at a certain temperature, in cells it almost always follows the harvesting step, often in pbs, at a certain pH.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Retained	**Definition:** To hold in tact. Just the word retained, not a biological term. **Why is it done:** In the few sentences you sent me it talks about using a special container to retain cell membranes, there is one that talks about looking at a compound in mass spec and then modifying it and looking at it again and seeing that its peak was retained. Basically it is heavily context dependent because it is just a word meaning remained the same. You might do it to anything. **How is it done:** Generally after some experimental change has occurred the word is used to let the reader know that the result is no change. Or that some experimental condition is being held constant.
Solubilized	**Definition:** To make a substance soluble or more soluble. **Why is it done:** To get something to be dissolved in solution. You might do it to: a solid or precipitate, crystallized something or a crystal, pellets, protein, steroids, samples, membranes, anything. **How is it done:** In a detergent/solute/solution, at a certain temperature in a certain buffer for a certain amount of time, stirring in a particular way, column, overnight.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Subcloned	**Definition:** To move DNA from one vector to another. Different from molecular cloning in that in molecular cloning once the sequence has been ligated into the target cloning/replication proceeds while in subcloning replication/cloning has already occurred and has been separated from its parent vector and is ligated into the recipient vector. **Why is it done:** To introduce new DNA into a vector for whatever reason: sequencing, increased expression, to prepare plasmid for insertion into something else (organism), to add something to the N-terminal or c-terminal end of an existing gene, to create fusion proteins. You might do it to: cDNA, DNA fragment, gene, RNA, coding region, spacer region, a plasmid, an oligonucleotide, a sequence, DNA library, a site or a region of a gene, promoter, terminator, a domain. **How is it done:** Using/at restriction sites (particular sites), ligation/ligases, in frame, in a vector, in a particular buffer, at a particular temperature, for a particular time, as fusion proteins, between two genetic elements, restriction enzyme techniques. Same sentence structure as molecular cloning as well [BLANK] was cloned INTO [BLANK] the first blank being the gene or sequence of interest and the second blank being the target plasmid/vector/site.
Subjected	**Definition:** To cause something to undergo or experience something (experiment/treatment). Fancy way of saying the experiment was carried out. **Why is it done:** To carry out an experiment or some action. You might do it to: any sample or subject of interest. **How is it done:** Any number of ways depending on what is being subjected to what.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Supplemented	**Definition:** Add an extra element or amount to. **Why is it done:** Often done to cell growth media (supplement with nutrients). You might do it to: media, cells, extracts, reactions, reactants, phenomena. **How is it done:** Often supplemented with serum (serum being the thing that is supplementing the cells/media), by adding something to something, or if in past tense means something that contains something "X-supplemented-X" where the first X is what was put into the second X.
Synthesized	**Definition:** To combine or produce something (by synthesis bringing together or combining). **Why is it done:** To create a functional piece of DNA or RNA from stock components or more generally to create some new chemical species by combining two thingsfor whatever reason. You might do it to: DNA, oligonucleotides, plasmids, RNA probes/RNA, a chemical. **How is it done:** The general formula of a synthesis reaction is A + B = AB, that forumla is meant to describe smaller chemicals/atoms interacting but things like nucleotides (DNA) or peptides (protein) are also synthesized as they are made by combining chains of nucleic acids or amino acids respectively. These synthesis reactions happen in naturally of course but are also used to create new nucleotides and peptides for study, the chemical reactions that occur have many steps. You can see the steps for DNA under Purine and Pyrimidine nucleotide synthesis sections. Protein of course is synthesized based on an mRNA transcript in a ribosome (which gets its sequence from being copied off the DNA) here is a video that explains pretty well. Protein folding is just as important as sequence as well, so you can have the correct sequence but a misfolded denatured useless protein. There are things in the cell which assist a newly synthesized protein in folding correctly[1].

Continued on next page

[1] https://www.youtube.com/watch?v=gG7uCskUOrA

Continued from previous page

Verb	Definition and their usage
Transduced	**Definition:** The transfer of genetic material from one organism to another (generally bacteria) via a genetic vector (especially a bacteriophage). **Why is it done:** In the one sentence provided it is being done to create a particular strain of E.coli. You might do it to: DNA, plasmid, viral vector, bacteriophage (a virus affects bacteria), DNA fragment, gene, sequence. **How is it done:** Using a particular bacteriophage, into a certain cell type, generalized transduction any bacterial gene is transferred, specialized transduction, a restricted set of bacterial genes are transferred.
Transfected	**Definition:** Actively introducing nucleic acids into eukaryotic cells (e.g., human cells). **Why is it done:** Do give a cell/cell culture new DNA for any reason Ex. Protein coding gene, selectable marker, etc...). You might do it to: any eukaryotic cell (not 'prokaryotic' prokaryotic cells (e.g., bacterial cells). Eukaryotic cells are those that contain a nucleus. **How is it done:** Many ways depending on the cell type being transfected, using a particular transfection reagent, a particular buffer, and with particular volumes of: Media, transfection reagent, and DNA. Usually followed by incubation, and follows splitting cells and incubation. Can be done using electroporation instead of transfection reagents. General sentence structure for determining method of transfection: cells were transfected with/using X (x will either say electroporation or some brand name chemical which will be the transfection reagent).

Continued on next page

Continued from previous page

Verb	Definition and their usage
Transferred	**Definition:** No specific biochemical meaning, just means to move from one place to another. **Why is it done:** Any time there is something in one location that should go to another. Some things in the cell are transferred naturally as part of biological processes, for example a cell membrane surface receptor can be transferred/transported to the nucleus via a nuclear localization signal and a vesicle. You might do it to: anything. **How is it done:** Any conceivable way of moving something, ex. Pouring, pipetting, bacterial conjugation of a plasmid could even be referred to as transferring the plasmid.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Transformed	**Definition:** Refers to any cell that has been genetically altered after the uptake of genetic material from outside the cell. So a cell that has been transfected or transduced will also have been transformed. But transformed is more often used to describe prokaryotic cells as people usually just say transfected cells for eukaryotic cells. Also humans cannot directly transduce a bacterium because in order for that term to apply a bacterium itself has to facilitate that gene transfer to another. So as a general summary of how you will usually see the terms used: Human adding DNA to prokaryote = transformed; Human adding DNA to eukaryote = transfected; Bacterium giving DNA to Bacterium (prokaryotes) **Why is it done:** To give a cell new genetic material, particularly a prokaryotic cell. You might do it to: any cell (particularly prokaryote). **How is it done:** Different methods depending on the cell type. For bacteria you need to make the cells "competent" such that they are permeable to the DNA entering them. The most common method of doing this is called "heat shock" where the bacteria are stored somewhere cold before being exposed to heat, which makes their membranes competent. Electroporation also works with bacteria as well as eukaryotic yeast and plants. Also in plants transformation can be accomplished with agrobacterium as a vector which infects the plant cells and sends and infects them with the DNA you loaded the bacterium with, or a gene gun where DNA is loaded onto a particle (often a gold particle) then fired at a high speed past the plant cell wall. Often followed by incubation at a certain temperature for bacteria and yeast.

Continued on next page

Continued from previous page

Verb	Definition and their usage
Washed	**Definition:** To remove unwanted species from a species of interest. **Why is it done:** Usually one of the last steps of the purification process for DNA and protein. You might do it to: DNA, protein, plasmid, molecule. Could be anything though. **How is it done:** Usually done multiple times during purification, with a certain wash buffer that doesnt solubilize the species of interest. Often done in a centrifuge. Washes of cells usually done with PBS (to rinse off media or wash cells off the plate entirely). The subject of interest is usually fixed in place somehow and the wash should not disrupt that, until it is time for the elution step (which follows washing in many cases) where the subject of interest can finally leave the place where it was fixed.

Table E.1: List of Common Procedural Verbs of Biochemistry Articles

Appendix F

List of Frames for Procedural Verbs

FRAMES for amplify-111
NP V NP NP
[NP Responses] [VP were amplified] [NP 10] , [ADVP 000-fold] ([NP Grass P511 High Performance AC Amplifier]).
Example: '' Responses were amplified 10,000-fold (Grass P511 High Performance AC Amplifier)"
Syntax: PATIENT V CONDITION INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V NP PP
[NP The 3′ UTR] [PP of] [NP the luciferase gene] [VP containing] [NP barcode sequences] [VP was amplified] ([NP Phusion high fidelity master mix] ; [NP New England Biolabs]) [PP from] [NP cellular RNA] ([NP cDNA]) or [NP DNA].
Example: '' The 3′ UTR of the luciferase gene containing barcode sequences was amplified (Phusion high fidelity master mix; New England Biolabs) from cellular RNA (cDNA) or DNA…"
Syntax: PATIENT V INSTRUMENT (+SRC) THEME
Semantics: MANNER (DURING(E), PATIENT)
NP V NP
[NP Array-generated oligos] [VP were amplified] ([NP four cycles]) .
Example: 'Array-generated oligos were amplified (four cycles).'
Syntax: PATIENT V REPETITION
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for anneal-112
NP V PP PP PP PP
[NP DNA primers] [VP were annealed] [PP to] [NP 3 µg] [PP of] [NP total RNA] [PP by] [VP incubating] [PP at] [NP 95 °C] [PP for] [NP 2 min].
Example: '' DNA primers were annealed to 3 µg of total RNA by incubating at 95 °C for 2 min "
Syntax: PATIENT V GOAL INSTRUMENT TEMP TIME
Semantics: MANNER (DURING(E), PATIENT)
NP V NP PP
[NP Two complimentary oligonucleotides] , [NP spd130 and spd133] , [VP were annealed to introduce] [NP a single Zif268 binding site] [PP in] [NP pFS414] [VP to form] [NP pFS410].
Example: "Two complimentary oligonucleotides, spd130 and spd133, were annealed to introduce a single Zif268 binding site in pFS414 to form pFS410."
Syntax: PATIENT V GOAL THEME
Semantics: MANNER(DURING(E), PATIENT)

FRAMES for biotinylate-113
NP V
[NP Full-length cDNAs] [VP were biotinylated].
Example: "Full-length cDNAs were biotinylated"
Syntax: PATIENT V
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for centrifuge-114
NP V PP PP PP
[NP The lysate] [VP was centrifuged] [PP at] [NP 13] , [NP 000 g] [PP for] [NP 10 min] [PP in] [NP a 1.5-ml Eppendorf tube] .
Example: '' The lysate was centrifuged at 13,000 g for 10 min in a 1.5-ml Eppendorf tube."
Syntax: PATIENT V CONDITION TIME CONDITION
Semantics: MOTION (DURING(E), PATIENT)
NP V PP PP PP
[NP The suspension] [VP was subsequently centrifuged] [PP at] [NP 100] , [NP 000 g] [PP for] [NP 1 h] [PP at] [NP 4 °C].
Example: "The suspension was subsequently centrifuged at 100,000 g for 1 h at 4 °C"
Syntax: PATIENT V CONDITION TIME TEMP
Semantics: MOTION (DURING(E), PATIENT)
PP NP V PP PP PP VP
[PP After] [NP sonication] , [NP the suspension] [VP was centrifuged] [PP at] [NP 48] , [NP 000 g] [PP at] [NP 4°C] [PP for] [NP 40 min] [VP to isolate] [NP inclusion bodies].
Example: '' After sonication, the suspension was centrifuged at 48,000 g at 4°C for 40 min to isolate inclusion bodies. "
Syntax: CONDITION PATIENT V CONDITION TEMP TIME GOAL
Semantics: MOTION (DURING(E), PATIENT)
NP V NP
[NP Samples] [VP were centrifuged] [VP to remove] [NP insoluble material].
Example: "Samples were centrifuged to remove insoluble material"
Syntax: PATIENT V GOAL
Semantics: MOTION (DURING(E), PATIENT)
NP V
[NP Cells] [VP were centrifuged]
Example: ''Cells were centrifuged"
Syntax: PATIENT V
Semantics: MOTION (DURING(E), PATIENT)

FRAMES for clone-115
NP V PP
[NP Array-generated oligos] [VP were cloned] [PP into] [NP a pRho-dsRED vector] ([NP Kwasnieski et al] . [NP 2012])
Example: '' Array-generated oligos were cloned into a pRho-dsRED vector (Kwasnieski et al. 2012)"
Syntax: PATIENT V GOAL
Semantics: MANNER (DURING(E), PATIENT)
PP NP V PP
[PP For] [NP the generation] [PP of] [NP lentiviral shRNA expression vectors] [VP targeting] [NP Denr] , [NP sequences] [PP from] [NP the TRC shRNA Library] [PP at] [NP the Broad Institute] [VP were cloned] [PP into] [NP pLKO.1puro backbone vector] ([NP Addgene] [INTJ no] . [NP 10878]) ([NP Moffat et al] . [NP 2006]) .
Example: "For the generation of lentiviral shRNA expression vectors targeting Denr, sequences from the TRC shRNA Library at the Broad Institute were cloned into pLKO.1puro backbone vector (Addgene no. 10878) (Moffat et al. 2006). "
Syntax: GOAL PATIENT V GOAL
Semantics: MANNER(DURING(E), PATIENT)
NP V PP VP
[NP ARS305] [VP was cloned] [PP into] [NP EagI- , MluI-cut pFS410] [VP to create] [NP pFS416] .
Example: "ARS305 was cloned into EagI-, MluI-cut pFS410 to create pFS416."
Syntax: PATIENT V GOAL GOAL
Semantics: MANNER (DURING(E), PATIENT)
NP V PP NP
[NP The U7Ub25.2540 construct] [VP was then cloned] [PP into] [NP the 3xHA pcDNA3.1] (+) [NP vector] [VP using] [NP BamHI and EcoRI restriction sites] .
Example: ''The U7Ub25.2540 construct was then cloned into the 3xHA pcDNA3.1 (+) vector using BamHI and EcoRI restriction sites. "
Syntax: PATIENT V GOAL INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for collect-116
NP V
[NP The supernatants] [VP were collected] .
Example: “The supernatants were collected.”
Syntax: THEME V
Semantics: MOTION (DURING(E), THEME)
NP V PP PP
[NP The supernatant] [VP was collected] [PP for] [NP immunoblotting] [PP after] [NP the tube] [VP was placed] [PP on] [NP a magnet] ([NP Fig] . [LST 4a]) .
Example: “The supernatant was collected for immunoblotting after the tube was placed on a magnet (Fig. 4a).”
Syntax: THEME V GOAL CONDITION
Semantics: MOTION (DURING(E), THEME)
NP V PP
[NP Forty] ([NP 30-kDa sample]) or [NP twenty-five] (10-kDa sample) [NP fractions] [VP were collected] [PP over] [NP a 70-min gradient] [VP beginning] [PP with] [NP 0.1 % acetic acid] [PP in] [NP 90 % acetonitrile] ([NP aq.]) and [VP ending] [PP with] [NP 0.1 % formic acid] [PP in] [NP 30 % acetonitrile] ([NP aq.]) .
Example: “Forty (30-kDa sample) or twenty-five (10-kDa sample) fractions were collected over a 70-min gradient beginning with 0.1% acetic acid in 90% acetonitrile (aq.) and ending with 0.1% formic acid in 30% acetonitrile (aq.).”
Syntax: THEME V TIME INSTRUMENT
Semantics: MOTION (DURING(E), THEME)
PP NP V PP
[PP For] [NP each sample] , [NP 400 ODs] [PP of] [NP cells] [VP were collected] [PP by] [NP centrifugation] .
Example: “For each sample, 400 ODs of cells were collected by centrifugation.”
Syntax: CONDITION THEME V INSTRUMENT
Semantics: MOTION (DURING(E), THEME)
NP V NP PP
[NP Samples] [VP were collected] [NP every 5 min] [PP between] [NP 10 and 80 min] .
Example: “Samples were collected every 5 min between 10 and 80 min.”
Syntax: THEME V REPETITION TIME
Semantics: MOTION (DURING(E), THEME)
NP V PP
[NP Two ODs] [PP of] [NP cells] [VP were collected] [PP for] [NP genomic DNA isolation] .
Example: “Two ODs of cells were collected for genomic DNA isolation.”
Syntax: THEME V GOAL
Semantics: MOTION (DURING(E), THEME)
NP V PP PP NP
[NP Vacuoles] [VP were collected] [PP by] [NP 5-min centrifugation] [PP at] [NP 4] , [NP 500 g] , [NP 2 °C] .
Example: “Vacuoles were collected by 5-min centrifugation at 4,500 g, 2 °C.”
Syntax: THEME V INSTRUMENT CONDITION TEMP
Semantics: MOTION (DURING(E), THEME)
NP V NP NP
[NP We] [VP collected] [NP all SNPs] [NP that] [VP intersect annotated] [NP Kozak regions] .
Example: “We collected all SNPs that intersect annotated Kozak regions.”
Syntax: AGENT V THEME CONDITION
Semantics: MOTION (DURING(E), THEME)

FRAMES for concentrate-117
NP V PP
[NP The combined organic extract] [VP were concentrated] [PP by] [NP rotary evaporation] .
Example: “The combined organic extract were concentrated by rotary evaporation.”
Syntax: PATIENT V INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V PP PP PP
[NP Elution fractions] [VP containing] [NP ligase protein] [VP were concentrated] [PP under] [NP high pressure] [PP in] [NP a stirred-cell concentrator unit] [PP with] [NP a 5] , [NP 000-MWCO Ultracel Ultrafiltration cellulose membrane] ([NP Millipore]) .
Example: “ Elution fractions containing ligase protein were concentrated under high pressure in a stirred-cell concentrator unit with a 5,000-MWCO Ultracel Ultrafiltration cellulose membrane (Millipore).”
Syntax: PATIENT V CONDITION CONDITION INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V
[NP The solution] [VP was concentrated] .
Example: “The solution was concentrated.”
Syntax: PATIENT V
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for conjugate-118
PP NP V PP NP VP
[PP Following] [NP Boc deprotection] and [NP ion exchange] , [NP the compound] ([NP 12]) [VP was conjugated] [PP to] [NP 6-] (([NP biotinoyl]) [NP amino]) [NP hexanoic acid] [VP using] [NP standard peptide] [VP coupling] [NP methods] [VP to yield] [NP the desired YM-1-biotin] ([NP 2] ; [NP YM-1-biotin]) [PP as] [NP a dark red solid] .
Example: “Following Boc deprotection and ion exchange, the compound (12) was conjugated to 6-((biotinoyl)amino)hexanoic acid using standard peptide coupling methods to yield the desired YM-1-biotin (2 ; YM-1-biotin) as a dark red solid.”
Syntax: CONDITION PATIENT V COFACTOR INSTRUMENT GOAL
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for denature-119
NP V PP PP PP
[NP 2 µg] [PP of] [NP genomic DNA] [VP denatured] [PP for] [NP 30 min] [PP at] [NP 37°C] [PP in] [NP NaOH 0.4 N]
Example: “ 2 µg of genomic DNA denatured for 30 min at 37°C in NaOH 0.4 N ”
Syntax: PATIENT V TIME TEMP INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V VP
[NP Libraries] [VP were denatured] [VP using] [NP NaOH] .
Example: “Libraries were denatured using NaOH.”
Syntax: THEME V INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for culture-120
NP V PP PP PP
[NP Amoebae] [VP used] [PP in] [NP these experiments] [VP were cultured] [PP in] [NP amoeba growth medium] [PP for] [ADVP no longer] [NP than 1 month] [PP after] [NP removal] [PP from] [NP mouse brain] .
Example: “Amoebae used in these experiments were cultured in amoeba growth medium for no longer than 1 month after removal from mouse brain.”
Syntax: THEME V INSTRUMENT TIME CONDITION
Semantics: MANNER (DURING(E), THEME)
NP V PP PP
[NP Mouse-passaged amoebae] [VP were cultured] [PP in] [NP Oxoid medium] [PP at] [NP 37 °C] .
Example: “ Mouse-passaged amoebae were cultured in Oxoid medium at 37 °C.”
Syntax: THEME V INSTRUMENT TEMP
Semantics: MANNER (DURING(E), THEME)
NP V PP VP PP VP NP
[NP CGR8 mESCs] [VP were cultured] [PP on] [NP gelatin-coated dishes] ([ADJP feeder-free] , [VP to avoid] [NP DNA contamination] [PP by] [NP MEF cells]) [PP in] [NP GMEM medium] [VP supplemented] [PP with] [NP 10 %] [PP of] [NP FBS] and [NP 1000 units/mL] [PP of] [NP LIF] .
Example: “CGR8 mESCs were cultured on gelatin-coated dishes (feeder-free, to avoid DNA contamination by MEF cells) in GMEM medium supplemented with 10% of FBS and 1000 units/mL of LIF.”
Syntax: THEME V CONDITION GOAL INSTRUMENT COFACTOR COFACTOR
Semantics: MANNER (DURING(E), THEME)
NP V PP
[NP Cells] [VP were cultured] [PP for] [NP up to 9 d] .
Example: “Cells were cultured for up to 9 d.”
Syntax: THEME V TIME
Semantics: MANNER (DURING(E), THEME)
NP V PP PP PP
[NP B] . [NP subtilis str] . [NP 168] ([VP kindly provided] [PP by] [NP A] . [NP Soma] [PP of] [NP Chiba University]) [VP was cultured] [PP in] [NP LB medium] [PP at] [NP 37 °C] [PP for] [NP 24 h] .
Example: “B. subtilis str. 168 (kindly provided by A. Soma of Chiba University) was cultured in LB medium at 37 °C for 24 h.”
Syntax: THEME V INSTRUMENT TEMP TIME
Semantics: MANNER (DURING(E), THEME)
NP V PP NP PP PP
[NP Saccharomyces cerevisiae BY4742] ([NP Euroscarf]) [VP was cultured] [PP in] [NP YPD] ([NP 1 % yeast extract] , [NP 2 % peptone] and [NP 2 % glucose]) or [NP YPG] ([NP 1 % yeast extract] , [NP 2 % peptone] and [NP 3 % glycerol]) [PP at] [NP 30 °C] [PP for] [NP 18 h] .
Example: “Saccharomyces cerevisiae BY4742 (Euroscarf) was cultured in YPD (1% yeast extract, 2% peptone and 2% glucose) or YPG (1% yeast extract, 2% peptone and 3% glycerol) at 30 °C for 18 h.”
Syntax: THEME V INSTRUMENT INSTRUMENT TEMP TIME
Semantics: MANNER (DURING(E), THEME)
NP V PP PP PP NP PP NP
[NP E] . [NP coli strain A19] [VP was cultured] [PP in] [NP LB medium] [PP at] [NP 37 °C] [PP for] [NP 18 h] ([NP stationary phase]) or [PP for] [NP 4-5 h] ([NP mid-log phase]) .
Example: “E. coli strain A19 was cultured in LB medium at 37 °C for 18 h (stationary phase) or for 4-5 h (mid-log phase).”
Syntax: THEME V INSTRUMENT TEMP TIME CONDITION TIME CONDITION
Semantics: MANNER (DURING(E), THEME)
NP V PP
[NP DCs] [VP were cultured] [PP on] [NP poly-L-lysine-coated coverslips] .
Example: “DCs were cultured on poly-L-lysine-coated coverslips.”
Syntax: THEME V INSTRUMENT
Semantics: MANNER (DURING(E), THEME)

FRAMES for dilute-121
NP V NP PP
[NP Primary antibodies] [VP were diluted] [NP 1] : [NP 50] [PP in] [VP blocking] [NP reagent] .
Example: “Primary antibodies were diluted 1:50 in blocking reagent.”
Syntax: PATIENT V CONDITION BUFFER
Semantics: MANNER (DURING(E), PATIENT)
NP V PP PP
[NP Purified kinase] ([NP 100 nM]) , [NP mammalian lysate] ([NP 0.5 mg/ml]) and [NP HT- 1] ([NP 500 nM]) [VP were diluted] [PP in] [NP PBS] [PP in] [NP a 96-well U-bottom plate] .
Example: “Purified kinase (100 nM), mammalian lysate (0.5 mg/ml) and HT- 1 (500 nM) were diluted in PBS in a 96-well U-bottom plate.”
Syntax: PATIENT V BUFFER INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V ADVP PP
[NP The supernatant] [VP was diluted] [ADVP six-fold] [PP in] [NP ice-cold PS buffer] .
Example: “The supernatant was diluted six-fold in ice-cold PS buffer.”
Syntax: GOAL PATIENT V INSTRUMENT CONDITION
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for elute-122
NP V ADVP PP PP
[NP The resulting supernatant] [VP was eluted] [ADVP stepwise] [PP with] [NP 10 %] , [NP 35 % and 80 %] ([NP v/v]) [NP acetonitrile] [PP in] [NP 0.08 %] ([NP v/v]) [NP trifluoroacetic acid] ([NP TFA]) .
Example: “ The resulting supernatant was eluted stepwise with 10%, 35% and 80% (v/v) acetonitrile in 0.08% (v/v) trifluoroacetic acid (TFA). ”
Syntax: PATIENT V CONDITION INSTRUMENT INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V PP
[NP The digested peptide mix] [VP was eluted] [PP with] [NP 1 : 1 acetonitrile /water] .
Example: “The digested peptide mix was eluted with 1:1 acetonitrile /water.”
Syntax: PATIENT V INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for equilibrate-123
NP V PP PP PP PP
[NP Samples] [VP equilibrated] [PP with] [NP 50 mM sodium acetate] , [NP pH 4.4] , [PP at] [NP a flow rate] [PP of] [NP 1 ml] [PP per] [NP min] [PP at] [NP 4°C] .
Example: “Samples equilibrated with 50 mM sodium acetate, pH 4.4, at a flow rate of 1 ml per min at 4°C.”
Syntax: THEME V COFACTOR CONDITION TIME TEMP
Semantics: MANNER (DURING(E), THEME)
NP V PP NP VP
[NP a Superdex-200 column] ([NP GE Healthcare]) [VP equilibrated] [PP with] [NP 10 mM NaPO 4 , 50 mM NaCl] and [NP 1 mM EDTA] ([NP pH 6.7]) [VP to remove] [NP aggregated protein] .
Example: “a Superdex-200 column (GE Healthcare) equilibrated with 10 mM NaPO 4, 50 mM NaCl and 1 mM EDTA (pH 6.7) to remove aggregated protein.”
Syntax: THEME V COFACTOR COFACTOR GOAL
Semantics: MANNER (DURING(E), THEME)

FRAMES for dilute-124
NP V NP PP
[NP Primary antibodies] [VP were diluted] [NP 1] : [NP 50] [PP in] [VP blocking] [NP reagent] .
Example: “Primary antibodies were diluted 1:50 in blocking reagent.”
Syntax: PATIENT V CONDITION BUFFER
Semantics: MANNER (DURING(E), PATIENT)
NP V PP PP
[NP Purified kinase] ([NP 100 nM]) , [NP mammalian lysate] ([NP 0.5 mg/ml]) and [NP HT- 1] ([NP 500 nM]) [VP were diluted] [PP in] [NP PBS] [PP in] [NP a 96-well U-bottom plate] .
Example: “Purified kinase (100 nM), mammalian lysate (0.5 mg/ml) and HT- 1 (500 nM) were diluted in PBS in a 96-well U-bottom plate.”
Syntax: PATIENT V BUFFER INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V ADVP PP
[NP The supernatant] [VP was diluted] [ADVP six-fold] [PP in] [NP ice-cold PS buffer] .
Example: “The supernatant was diluted six-fold in ice-cold PS buffer.”
Syntax: PATIENT V CONDITION INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)

FRAMES for wash -125
NP V NP PP PP
[NP Slides] [VP were washed] [NP three times] [PP for] [NP 10 min] [PP in] [NP PBS] .
Example: “Slides were washed three times for 10 min in PBS.”
Syntax: PATIENT V REPETITION TIME INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V PP NP PP NP PP
[NP The cells] [VP were then washed] [PP with] [NP medium] ([NP three times] [PP for] [NP 5 min] [NP each] [PP at] [NP 37 °C]) .
Example: “The cells were then washed with medium (three times for 5 min each at 37 °C).”
Syntax: PATIENT V INSTRUMENT REPETITION TIME CONDITION TEMP
Semantics: MANNER (DURING(E), PATIENT)
NP V ADVP NP PP
[NP The resin] [VP was washed] ([ADVP twice] , [NP ten-bed volumes]) [PP with] [NP immobilization buffer] .
Example: “The resin was washed (twice, ten-bed volumes) with immobilization buffer.”
Syntax: PATIENT V REPETITION CONDITION INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V ADVP PP PP
[NP The pellet] [VP was washed] [ADVP twice] [PP before] [NP resuspension] [PP in] [NP the same buffer] .
Example: “The pellet was washed twice before resuspension in the same buffer.”
Syntax: PATIENT V REPETITION CONDITION INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)

Appendix G

XML file for Annotation of Semantic Roles and Rhetorical Moves

```
<?xml version='1.0' encoding='UTF-8'?>
<GateDocument version="3">
<!-- The document's features-->

<GateDocumentFeatures>
<Feature>
  <Name className="java.lang.String">gate.SourceURL</Name>
  <Value className="java.lang.String">file:/Users/anna/Documents/GATE%20files/Example1
.txt</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">MimeType</Name>
  <Value className="java.lang.String">text/plain</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">docNewLineType</Name>
  <Value className="java.lang.String">CRLF</Value>
</Feature>
</GateDocumentFeatures>
<!-- The document content area with serialized nodes -->

<TextWithNodes><Node id="0"/>T<Node id="1"/>he <Node id="4"/>ca. 900 bp PCR products<N
ode id="27"/> <Node id="28"/>were digested<Node id="41"/> <Node id="42"/>with NdeI and
 HindIII<Node id="63"/> and <Node id="68"/>ligated<Node id="75"/> <Node id="76"/>into
pUC19<Node id="86"/>.&#xd;
<Node id="89"/>&#xd;
<Node id="91"/>Steady-state kinetics constants, Km and kcat<Node id="135"/>, <Node id=
"137"/>were determined<Node id="152"/> <Node id="153"/>by fitting initial velocity ver
sus substrate concentration data directly to the Michaelis equation using CurveFit [36
<Node id="270"/>]<Node id="271"/>.<Node id="272"/></TextWithNodes>
<!-- The default annotation set -->

<AnnotationSet>
</AnnotationSet>

<!-- Named annotation set -->

<AnnotationSet Name="Original markups">
<Annotation Id="0" Type="paragraph" StartNode="0" EndNode="89">
</Annotation>
<Annotation Id="1" Type="paragraph" StartNode="91" EndNode="272">
</Annotation>
<Annotation Id="3" Type="Predicate" StartNode="28" EndNode="41">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Verb</Value>
</Feature>
</Annotation>
<Annotation Id="4" Type="SemanticRole" StartNode="4" EndNode="27">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Patient</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">3</Value>
</Feature>
</Annotation>
<Annotation Id="5" Type="Semantic_Role:Instrument" StartNode="42" EndNode="63">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Catalyst</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
```

```
  <Value className="java.lang.Integer">3</Value>
</Feature>
</Annotation>
<Annotation Id="6" Type="Predicate" StartNode="68" EndNode="75">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Verb</Value>
</Feature>
</Annotation>
<Annotation Id="7" Type="SemanticRole" StartNode="76" EndNode="86">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Goal:Physical</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">6</Value>
</Feature>
</Annotation>
<Annotation Id="8" Type="ArgMoves" StartNode="0" EndNode="63">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Description_method</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">3</Value>
</Feature>
</Annotation>
<Annotation Id="9" Type="ArgMoves" StartNode="1" EndNode="86">
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">6</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Description_method</Value>
</Feature>
</Annotation>
<Annotation Id="10" Type="SemanticRole" StartNode="4" EndNode="27">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Patient</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">6</Value>
</Feature>
</Annotation>
<Annotation Id="11" Type="Predicate" StartNode="137" EndNode="152">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Verb</Value>
</Feature>
</Annotation>
<Annotation Id="12" Type="SemanticRole" StartNode="91" EndNode="135">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Patient</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">11</Value>
</Feature>
</Annotation>
```

```
<Annotation Id="13" Type="Semantic_Role:Instrument" StartNode="153" EndNode="271">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Mathematical</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">11</Value>
</Feature>
</Annotation>
<Annotation Id="14" Type="ArgMoves" StartNode="91" EndNode="270">
<Feature>
  <Name className="java.lang.String">Type</Name>
  <Value className="java.lang.String">Reference_to_method</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Relation</Name>
  <Value className="java.lang.Integer">11</Value>
</Feature>
</Annotation>
</AnnotationSet>

</GateDocument>
```

Appendix H

A Sample of Complete Article from Our Dataset [46]

ELSEVIER

FEMS Microbiology Letters 201 (2001) 229–235

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

In vitro reconstruction of the biosynthetic pathway of peptidoglycan cytoplasmic precursor in *Pseudomonas aeruginosa*

Ahmed El Zoeiby [a], François Sanschagrin [a], Pierre C. Havugimana [b], Alain Garnier [b], Roger C. Levesque [a,*]

[a] *Centre de Recherche sur la Fonction, Structure et Ingénierie des Protéines, Faculté de Médecine, Pavillon Charles-Eugène Marchand, Ste-Foy, QC, Canada G1K 7P4*

[b] *Département de Génie chimique, Faculté des Sciences et Génie, Université Laval, Ste-Foy, QC, Canada G1K 7P4*

Received 30 April 2001; received in revised form 29 May 2001; accepted 30 May 2001

First published online 25 June 2001

Abstract

Bacterial peptidoglycan is the cell wall component responsible for maintaining cell integrity against osmotic pressure. Biosynthesis of the cytoplasmic precursor UDP-*N*-acetylmuramyl pentapeptide is catalyzed by the Mur enzymes. Genomic analysis of the three regions encoding Mur proteins was achieved. We have cloned and over-expressed the *murA*, *-B*, *-D*, *-E* and *-F* genes of *Pseudomonas aeruginosa* in pET expression system by adding a His–Tag to the C-termini of the proteins. Mur proteins were purified to homogeneity by a single chromatographic step on affinity nickel columns. Protein identities were verified through N-terminal sequencing. Enzyme activity was proved by the identification of the pathway's final product. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.

Keywords: Cell wall biosynthesis gene; Bacterial cell wall; Peptidoglycan; Mur enzyme

1. Introduction

Bacterial cell wall polymer, peptidoglycan, is essential for cell survival by maintaining cell integrity against osmotic pressure [1]. Furthermore, peptidoglycan is a unique bacterial structure absent in eukaryotic host cells and it represents a potential antimicrobial target. The biosynthetic pathway of peptidoglycan is a complex two-stage process. The first stage occurs in the cytoplasm and it consists of the formation of the monomeric building block *N*-acetylglucosamine-*N*-acetylmuramyl pentapeptide. The first committed step in the pathway is the condensation of phospho(enol)pyruvate (PEP) and UDP-*N*-acetylglucosamine in a reaction catalyzed by MurA. This is followed by a MurB-catalyzed reduction of the enol-pyruvate moiety to D-lactate, yielding UDP-*N*-acetylmuramate. A series of ATP-dependent amino acid ligases (MurC, MurD, MurE and MurF) catalyze the stepwise addition of the pentapeptide side chain on the newly reduced D-lactyl group. The second stage involves the transfer of the precursor across the membrane by a lipophilic carrier and its addition to the growing cell wall polymer by the enzymatic action of penicillin-binding proteins (PBPs) [2]. Many antibiotics in clinical use, mostly β-lactams and glycopeptides, interfere with the action of PBPs. However, all the enzymes involved in the early cytoplasmic steps of the pathway are not inhibited by known antibiotics or synthetic chemicals, except for MurA, which is inhibited by phosphonomycin [3].

The amino acid ligases are essential enzymes, highly specific and they occur only in eubacteria, thus they represent targets of particular interest. These enzymes contain highly conserved regions [4–6] and they operate via a similar mechanism involving carboxyl activation of the nucleotide substrate to an acylphosphate intermediate; and fol-

* Corresponding author. Tel.: +1 (418) 656-3070; Fax: +1 (418) 656-7176.
E-mail address: rclevesq@rsvs.ulaval.ca (R.C. Levesque).

Abbreviations: DMSO, dimethylsulfoxide; IPTG, isopropyl β-D-thiogalactopyranoside; PBP, penicillin-binding protein; PCR, polymerase chain reaction; PEP, phospho(enol)pyruvate; PVDF, polyvinylidene difluoride; SDS–PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis

0378-1097/01/$20.00 © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.
PII: S0378-1097(01)00268-3

lowed by nucleophilic attack by the amino group of the condensing amino acids, with the formation of a peptide bond and the elimination of a phosphate group [7–12]. Inhibitors have been designed for MurD [9,13–15] and MurE [16,17]. Mechanistic and structural studies of the Mur enzymes and screening of these enzymes for inhibitors are severely hampered because of the lack of pathway intermediates [18]. The only commercially available substrate is UDP-*N*-acetylglucosamine, the substrate for MurA.

In this paper, we present the genomic analysis and organization of *murA* to *murF* genes in three distinct loci in the 6.3-Mb genome of *Pseudomonas aeruginosa*. We also present the purification of MurA, -B, -D, -E and -F in mg quantities at 99% homogeneity or more and the reconstruction of the enzymatic pathway for the biosynthesis of the pentapeptide peptidoglycan precursor. We reported the cloning and over-expression of *P. aeruginosa murC* elsewhere [6]. MurA to -F were combinatorially used to reconstruct the whole pathway in vitro and the final product was identified.

2. Materials and methods

2.1. General

All reagents were purchased from Sigma Aldrich (Oakville, ON, Canada) unless otherwise indicated. Buffer D was 20 mM potassium phosphate, 1 mM 2-mercaptoethanol, 0.1 mM $MgCl_2$, 15% (v/v) glycerol, pH 7.0 [19].

2.2. DNA manipulations, reagents and techniques

Restriction endonuclease and T4 ligase were obtained from New England Biolabs (Beverly, MA, USA). Agarose gel electrophoresis and plasmid DNA preparations were performed according to published procedures [20]. Recombinant plasmids containing *P. aeruginosa mur* genes were propagated in *Escherichia coli* NovaBlue (Novagen, Madison, WI, USA) prior to protein synthesis in *E. coli* BL21(λDE3) (Novagen).

2.3. Cloning of P. aeruginosa murA, –B, -D, -E and -F

Polymerase chain reaction (PCR) cloning was used to obtain MurA, -B, -D, -E and -F proteins with a His–Tag at their C-terminal. Upper and lower primers designed to contain appropriate restriction sites were designed as shown in Table 1. Five PCR reactions were performed with the upper and lower primers for each gene using genomic DNA of *P. aeruginosa* strain PAO1293 as the template. PCR conditions were optimized as follows: 30 cycles, denaturation at 95°C for 60 s, annealing at 55°C for 60 s, and extension at 72°C for 90 s, primers at 0.1 μM each, dNTPs (Amersham Pharmacia Biotech, Piscataway, NJ, USA) at 0.2 mM each, $MgCl_2$ at 2 mM, 5% dimethylsulfoxide (DMSO) in a final volume of 50 μl and adding 2.6 units of Expand high fidelity polymerase (Roche Diagnostics, Laval, QC, Canada) after Hot start of 7 min at 95°C. PCR products were purified using Qiaquick PCR purification kit (Qiagen, Chatsworth, CA, USA). Purified PCR products were digested with the restriction enzymes included in upper and lower primers and were cloned into the corresponding sites of the expression vectors pET30a and pET21 (Novagen) under the control of the bacteriophage T7 promoter.

2.4. DNA sequencing and computer analysis

Genomic analysis was done using data from the complete *P. aeruginosa* strain PAO1 sequence (www.pseudomonas.com) [21]. The sequences reported have the GenBank accession number AE004859 (*murA*), AE004723 (*murB*), AF110740 (*murC*), AY008276 (*murD*, *-E* and *-F*) and AE004091 (the complete genome). The DNA inserts in recombinant plasmids pMON3005, pMON3006, pMON3013, pMON3014 and pMON3009 (Table 1) were sequenced using T7 promoter primer and T7 terminator primer (Novagen). Sequence analyses were performed by the programs of Wisconsin Package Version 10.1, Genetics Computer Group (GCG), Madison, WI, USA.

2.5. Overproduction of P. aeruginosa MurA, -B, -D, -E and -F

The recombinant plasmids pMON3005, pMON3006, pMON3013, pMON3014 and pMON3009 (Table 1) were introduced into the *E. coli* host strain BL21(λDE3) (Novagen) by electroporation for expression of MurA, -B, -D, -E and -F respectively, with a His–Tag at their C-terminal. Overproduction was tested at two different incubation temperatures: 30 and 37°C, for three incubation periods: 3 h, 6 h and overnight (starting from the addition of isopropyl β-D-thiogalactopyranoside (IPTG)), using two different culture media: terrific broth and LB broth, and adding IPTG to two final concentrations: 0.5 mM and 1 mM (added after a cell density of $OD_{600\ nm} = 0.5$ was reached). Maximum protein yields in the soluble fractions were obtained after incubation for 6 h at 37°C using LB broth and adding IPTG to a final concentration of 1 mM. A small-scale overproduction pilot experiment using the optimized conditions showed that the soluble fractions of the proteins constitute 10%, 5%, 20%, 20% and 50% of their total protein fractions, respectively. Cultures were grown at 37°C in 1 l of LB broth containing 50 mg ml^{-1} kanamycin for MurA, -B and -F and 100 mg ml^{-1} ampicillin for MurD and -E, until a cell density of $OD_{600\ nm} = 0.5$ was reached. Cells were pelleted and resuspended in LB broth containing 1 mM IPTG. Cells were induced for 6 h, pelleted at $3000 \times g$ and frozen at −80°C [22].

2.6. Purification of recombinant P. aeruginosa Mur proteins

The cell pellet from 1 l of each induced culture was resuspended in 100 ml binding buffer consisting of 5 mM imidazole, 0.5 M NaCl and 20 mM Tris–HCl pH 7.9 (Novagen) and cells were disrupted by three passages through a French press. Cell debris were pelleted by centrifugation at 17 000 rpm using a Sorvall SA-600 rotor (41 837×*g*) for 30 min. The supernatant was loaded into a column containing 2.5 ml of freshly prepared Novagen resin. The column was washed with 25 ml binding buffer, followed by 15 ml wash buffer consisting of 60 mM imidazole, 0.5 M NaCl and 20 mM Tris–HCl pH 7.9 (Novagen). The protein was eluted with 3×2.5 ml elute buffer consisting of 250 mM imidazole, 0.5 M NaCl and 20 mM Tris–HCl pH 7.9. Each elute fraction was collected in a 15-ml tube containing 10 ml buffer D to dilute the imidazole preventing protein precipitation. Elute fractions were analyzed on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) and the fractions containing the purified protein were pooled and concentrated on an Amicon YM 10 membrane (Millipore Corporation, Bedford, MA, USA). The concentration of each purified protein was determined by the Bio-Rad protein assay (Bio-Rad Laboratories, Hercules, CA, USA).

2.7. Protein sequencing

Mur proteins were resolved on 10% SDS–PAGE, transferred to a polyvinylidene difluoride (PVDF) membrane (Bio-Rad Laboratories) in 10 mM CAPS buffer pH 11 containing 10% methanol. The bands corresponding in size to the purified proteins were identified by staining with Ponceau S and were subjected to N-terminal sequencing by automatic Edman degradation performed on an Applied Biosystems model 473A pulsed liquid protein sequencer.

2.8. Enzyme assay

MurA to -F enzymes were tested for activity by a one-pot assay, supplying UDP-*N*-acetylglucosamine, PEP, ATP, NADPH and amino acids. The pathway assay contained in a final volume of 500 μl: bis-tris propane (50 mM, pH 8.0), L-alanine, D-glutamate, *meso*-diaminopimelate, D-alanyl-D-alanine (1 mM each), UDP-*N*-acetylglucosamine (120 μM), PEP (120 μM), NADPH (250 μM), DTT (500 μM), $(NH_4)_2SO_4$ (25 mM), KCl (5 mM), $MgCl_2$ (5 mM), ATP (5 mM), MurA, MurB, MurC, MurD, MurE and MurF at a final concentration of 0.3 μM each (modified from [23] as follows: amino acids, UDP-*N*-acetylglucosamine, PEP, NADPH, ATP and enzymes were used in different concentrations, use of non-radioactive UDP-*N*-acetylglucosamine and different final volume). The reaction was allowed to proceed for 5 h at 37°C. The reaction mixture was filtered through an Amicon YM 30

Table 1
Oligonucleotide sequences containing putative restriction sites for cloning in pET vector and expected PCR product sizes for each *mur* gene amplified[a]

Gene	Upper primer	Restriction site	Lower primer	Restriction site	PCR product	Recombinant plasmid	Reference
murA	5′-GGGCGACATATGCAAATTTCCAACTGA-3′	*Nde*I	5′-ACAGCCTCCGCGAGCTCGGTAGG-3′	*Sac*I	1471 bp	pMON3005	this work
murB	5′-GCGCATATGAGCCTGGAACTGCAGAA-3′	*Nde*I	5′-CCAGCCGCACCCTGAGCTCAGATTG-3′	*Sac*I	1038 bp	pMON3006	this work
murC	5′-CGGACGGTGGTCGATGCCTGCCTG-3′	*Nde*I	5′-GGTTCATGCGCCCTTCCCTCCCT-3′	*Sac*I	1480 bp	pMON3004	[6]
murD	5′-GCTGCGTTGAGGACGAAGAGTTAATGAG-3′	*Mxe*I	5′-CCGACAGCATCACTCCGAGCTCCTCTAC-3′	*Xho*I	1379 bp	pMON3013	this work
murE	5′-GCGTGGCATATGCCTATGAGCCTGAACC-3′	*Nde*I	5′-CGAAGAGGCTCACTCGAGCGGCACCC-3′	*Xho*I	1482 bp	pMON3014	this work
murF	5′-GGGAGGTGCCATATGCTTGAGCCTCTTC-3′	*Nde*I	5′-CGGCCAGCAGGAGGAGCTCTGAC-3′	*Sac*I	1407 bp	pMON3009	this work

[a]Parameters for PCR amplification (see Section 2.3).

membrane to remove the enzymes before HPLC analysis. A control reaction was performed by adding all the substrates and cofactors and omitting the enzymes. The control was incubated for 5 h at 37°C then analyzed by HPLC.

2.9. LC/MS system

The sample separation and analysis were performed on a HPLC coupled to a mass spectrometer (LC/MS) (Agilent Technologies, Mississauga, ON, Canada, model HP 1100 LC-MSD) comprised of a quaternary pump, a vacuum degasser, a refrigerated autosampler, a column compartment, a variable wavelength detector and an electrospray ionization (ESI), quadrupole mass spectrometer detector (MSD). The system control and data evaluation were done on a HP ChemStation for LC/MS. Separation was done using a 10-μm particle size MonoQ anion-exchange column 10×10 cm (Amersham Pharmacia Biotech) at room temperature. Flow rate was 0.3 ml min^{-1}. Injection volume was 100 μl. An isocratic feed of 0.02 M NH_4OAc, pH 9.0, 30 min, followed by a linear gradient from 0.02 to 1.0 M NH_4OAc, pH 9.0 over 2 h were used (modified from [24]: use of a linear instead of a non-linear gradient for elution). Mass spectrometry detection was performed with the ESI set at V_{cap} = 4500 V, nebulizing gas pressure = 35 psi, drying gas flow rate = 13 l min^{-1}, drying gas temperature = 350°C, with the quadrupole scanning from 1180 to 1200 *m*/*z* every 1.03 s with a step size of 0.15 amu.

3. Results and discussion

3.1. Genomic analysis of cell wall biosynthesis loci

A close inspection and analysis of the complete PAO1 6.3-Mb *P. aeruginosa* sequence revealed three loci encoding the MurA to -F enzymes (Fig. 1). The *murA* gene is located at 4.98 Mb from the start (www.pseudomonas.com). The *murB* gene is located at 3.34 Mb, 1.64 Mb upstream of *murA*. The *murC*, *-D*, *-E* and *-F* genes are present in the *mra* cluster of cell division and cell wall biosynthesis genes. This cluster is located at 4.94 Mb, only 40 kb upstream of *murA*. The following tandem of genes of the *mra* cluster is transcribed from the same DNA strand in the same orientation: *ftsL*, *pbp3*, *murE*, *murF*, *mraY*, *murD*, *ftsW*, *murG*, *murC*, *ddlB*, *ftsQ*, *ftsA*, *ftsZ* and *envA*, suggesting that these genes may constitute a single operon. This region has exactly the same gene organiza-

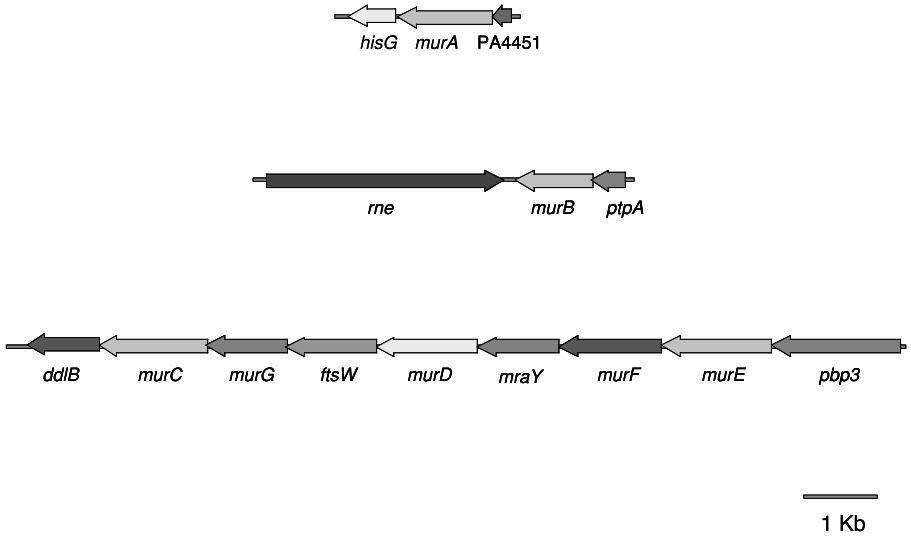

Fig. 1. *murA*, *murB* and *murC* to *-F* loci in *P. aeruginosa* genome. The three loci are drawn to scale using Redasoft Visual Cloning 2000. Gene organization and orientation is represented by arrows. Corresponding proteins are as follows: *hisG*, ATP-phosphoribosyl transferase; *murA*, UDP-*N*-acetylglucosamine(enol)pyruvate transferase; PA4451, conserved hypothetical protein; *rne*, ribonuclease E; *murB*, UDP-*N*-acetylpyruvylglucosamine reductase; *ptpA*, phosphotyrosine protein phosphatase; *ddlB*, D-alanine-{{INF:START}}$_{TEXTSPACE0}${{INF:END}}-alanine ligase; *murC*, UDP-*N*-acetylmuramate: L-alanine ligase; *murG*, UDP-*N*-acetylglucosamine: *N*-acetylmuramyl pentapeptide pyrophosphoryl-undecaprenol-*N*-acetylglucosamine transferase; *ftsW*, cell division protein FtsW; *murD*, UDP-*N*-acetylmuramyl-L-alanine: $_D$-glutamate ligase; *mraY*, phospho-*N*-acetylmuramyl pentapeptide transferase; *murF*, UDP-*N*-acetylmuramyl-$_L$-alanyl-D-glutamyl-2,6-diaminopimelate: $_D$-alanyl-$_D$-alanine ligase; *murE*, UDP-*N*-acetylmuramyl-$_L$-alanyl-$_D$-glutamate: 2,6-diaminopimelate ligase; *pbp3*, PBP 3.

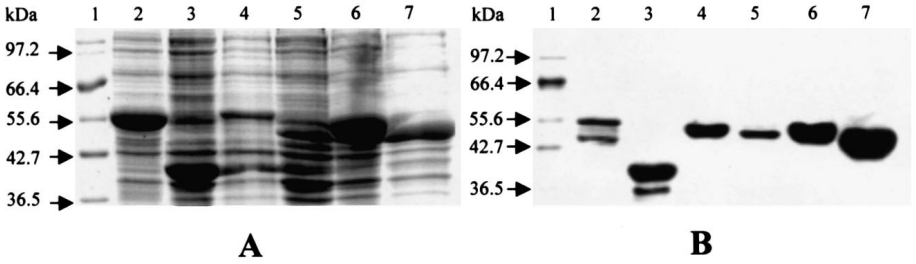

Fig. 2. A: Induced cell lysates analyzed on SDS–PAGE. Lane 1, protein marker broad range; lane 2, *E. coli* strain BL21+pMON3005; lane 3, *E. coli* strain BL21+pMON3006; lane 4, *E. coli* strain BL21+pMON3004; lane 5, *E. coli* strain BL21+pMON3013; lane 6, *E. coli* strain BL21+pMON3014; lane 7, *E. coli* strain BL21+pMON3009. B: Purified MurA to -F proteins analyzed on SDS–PAGE. Lane 1, protein marker broad range (New England Biolabs); lane 2, MurA (55 kDa); lane 3, MurB (39 kDa); lane 4, MurC (55 kDa); lane 5, MurD (51 kDa); lane 6, MurE (55 kDa); lane 7, MurF (52 kDa).

tion as the *mra* cluster in the 2-min region of the *E. coli* chromosome.

3.2. Cloning and sequencing of P. aeruginosa murA, -B, -D, -E and -F

Genomic DNA preparation of *P. aeruginosa* PAO1293 was used as a template for the PCRs used in the cloning of the *mur* genes separately using the primers designed to introduce the appropriate restriction sites at the ends of the PCR products. PCR products obtained corresponded to the length of each gene as shown in Table 1. Each of the cloned *mur* genes was sequenced in pET vector and the obtained sequences showed complete identity with *P. aeruginosa* corresponding genes in *Pseudomonas* Genome Project (www.pseudomonas.com). The recombinant plasmids encode recombinant Mur proteins with six His–Tag fusions to their C-termini. The advantage of the His–Tag fusion is to allow rapid purification of the protein by a single chromatographic step on an affinity nickel column.

3.3. Overproduction and purification of P. aeruginosa MurA, -B, -D, -E and -F

The recombinant plasmids pMON3005, pMON3006, pMON3013, pMON3014 and pMON3009 were grown in *E. coli* and cultures were induced with IPTG. The five transformants synthesized inducible proteins of the expected sizes of *P. aeruginosa* MurA, -B, -D, -E and -F; 55 kDa, 39 kDa, 51 kDa, 55 kDa and 52 kDa respectively. The ratio of the quantity of each protein in the soluble fraction was assessed versus its quantity in the total protein fraction. The estimation of the percentage of solubility of each protein was done by comparing the band brightness on SDS–PAGE (data not shown). Incubation temperature, incubation period, choice of culture medium and IPTG concentration were optimized for maximum protein yield (see Section 2.5). Each protein was purified in mg quantities to 99% homogeneity or more (Fig. 2).

3.4. N-terminal sequencing of P. aeruginosa MurA, -B, -C, -D, -E and -F

N-terminal sequencing of the first 15 amino acid residues of each purified protein including MurC [6] confirmed the identity of each. For MurA (Fig. 2A), the upper band at 55 kDa corresponded to the full-length protein and the lower band at 51 kDa showed a major sequence, LSPRGIIAMDKLIIT and a minor sequence, MDKLIITGGNRLDGE (residues in common are underlined) which are truncated MurA proteins lacking 54 and 62 N-terminal residues respectively. These two truncated species were co-purified with MurA due to their intact C-terminal with a His–Tag. For MurB (Fig. 2A), the upper band at 39 kDa corresponded to the full-length protein and the lower minor band at 36 kDa showed one major sequence, MKVAKDLVVSL. A standard protein–protein BLAST (blastp) search on PIR database showed 100% identity with the first 11 amino acids of SlyD protein (probable fkbP-type peptidyl-prolyl cis-trans isomerase) of *E. coli* which is a histidine-rich and a metal-binding protein initially detected as a persistent contaminant in immobilized metal affinity chromatography of recombinant proteins in *E. coli* [25]. The most probable reason of the observation of the SlyD contaminant only with purified MurB is the lower level of over-expression of MurB compared to the other Mur proteins (see Section 2.5) and hence the need for more concentration of the pooled elute fractions containing the purified protein, leading to the appearance of the minor contaminant. The first amino acid Met was absent in MurB, MurD and MurE.

3.5. Reconstruction of the biosynthetic pathway

MurA to -F were assayed simultaneously by reconstructing the murein pathway in vitro starting with the substrate for MurA, obviating the need and effort of preparing and purifying the substrates for the other enzymes

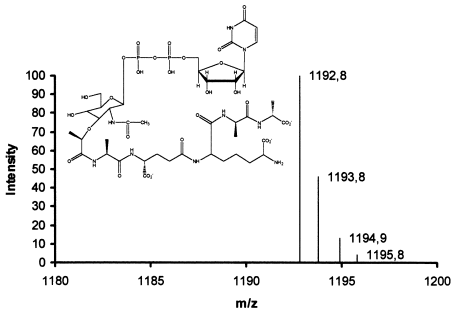

Fig. 3. Mass spectrum and chemical structure of the pathway final product (MurF product) UDP-*N*-acetylmuramyl pentapeptide eluted at 121 min on HPLC. The major peak *m/z* 1192.8 corresponds to the mass of the ionized pentapeptide after the loss of one proton.

which are not commercially available. The reaction mixture was analyzed by LC-MS. The pathway final product (product of MurF) UDP-*N*-acetylmuramyl pentapeptide was eluted after 121 min and its identity was confirmed by mass spectrometry, *m/z* 1192.8 (Fig. 3) corresponding to the calculated mass for the pentapeptide after the loss of one proton: $C_{41}H_{64}N_9O_{28}P_2$. Residual amounts of UDP-*N*-acetylmuramate (product of MurB) and UDP-*N*-acetylmuramyl tripeptide (product of MurE) were also detected at 108 and 158 min respectively, their identities being revealed by mass spectrometry. The control reaction including all reaction substrates and cofactors but lacking the enzymes was also analyzed by LC-MS. Only the starting substrate of the pathway, UDP-*N*-acetylglucosamine ($C_{17}H_{25}N_3O_{17}P_2Na_2$) was detected after 102 min, with *m/z* 606.3 corresponding to its calculated mass after the loss of two Na^+ and the gain of one proton: $C_{17}H_{24}N_3O_{17}P_2$. These results confirm the activity of the six purified Mur enzymes and the success in emulating the cytoplasmic steps of the biosynthesis of the peptidoglycan precursor in a single-pot assay using the tools in hand. Since the murein biosynthetic pathway comprises many validated antibacterial targets, the in vitro reconstruction of the cytoplasmic steps of this pathway is a tool of great value for antimicrobial drug discovery. This work would permit the development of novel screening tests for cell wall inhibitors. Such development of new strategies for antibiotic discovery has become highly imperative due to the alarming increase of bacterial resistance to all clinically useful antimicrobial agents.

Acknowledgements

This work was funded by the Canadian Bacterial Diseases Network via the Canadian Centers of Excellence and by a FCAR team grant. R.C.L. is a scholar of exceptional merit from Le Fonds de Recherche en santé du Québec, A.E.Z. obtained a studentship from Univ. Laval and from La Fondation Marc Bourgie. The authors thank Le Service de séquence de peptides de l'Est du Québec.

References

[1] Rogers, H.J., Perkins, H.R. and Ward, J.B. (1980) Biosynthesis of peptidoglycan. In: Microbial Cell Walls and Membranes (Rogers, H.J., Ed.), pp. 239–297. Chapman and Hall, London.
[2] van Heijenoort, J. (1994) Biosynthesis of the bacterial peptidoglycan unit. In: Bacterial Cell Wall (Ghuysen, J.-M. and Hakenbeck, R., Eds.), pp. 39–54. Elsevier, Amsterdam.
[3] Christensen, B.G., Leanza, W.J., Beatties, T.R., Patchett, A.A., Arison, B.H., Ormond, R.E., Kuehl Jr., F.A., Albers-Schonberg, G. and Jardetzky, O. (1969) Phosphonomycin: structure and synthesis. Science 166, 123–125.
[4] Bouhss, A., Mengin-Lecreulx, D., Blanot, D., van Heijenoort, J. and Parquet, C. (1997) Invariant amino acids in the Mur peptide synthetases of bacterial peptidoglycan synthesis and their modification by site-directed mutagenesis in the UDP-MurNAc: L-alanine ligase from *Escherichia coli*. Biochemistry 36, 11556–11563.
[5] Eveland, S.S., Pompliano, D.L. and Anderson, M.S. (1997) Conditionally lethal *Escherichia coli* murein mutants contain point defects that map to regions conserved among murein and folyl poly-γ-glutamate ligases: Identification of a ligase superfamily. Biochemistry 36, 6223–6229.
[6] El Zoeiby, A., Sanschagrin, F., Lamoureux, J., Darveau, A. and Levesque, R.C. (2000) Cloning, over-expression and purification of *Pseudomonas aeruginosa murC* encoding uridine diphosphate *N*-acetylmuramate:L-alanine ligase. FEMS Microbiol. Lett. 183, 281–288.
[7] Falk, P.J., Ervin, K.M., Volk, K.S. and Ho, H.T. (1996) Biochemical evidence for the formation of a covalent acyl-phosphate linkage between UDP-*N*-acetylmuramate and ATP in the *Escherichia coli* UDP-*N*-acetylmuramate: L-alanine ligase-catalyzed reaction. Biochemistry 35, 1417–1422.
[8] Liger, D., Masson, A., Blanot, D., van Heijenoort, J. and Parquet, C. (1996) Study of the overproduced uridine-diphosphate-*N*-acetylmuramate: L-alanine ligase from *Escherichia coli*. Microb. Drug Resist. 2, 25–27.
[9] Tanner, M., Vaganay, S., van Heijenoort, J. and Blanot, D. (1996) Phosphinate inhibitors of the D-glutamic acid-adding enzyme of peptidoglycan biosynthesis. J. Org. Chem. 61, 1756–1760.
[10] Vaganay, S., Tanner, M., van Heijenoort, J. and Blanot, D. (1996) Study of the reaction mechanism of the D-glutamic acid-adding enzyme from *Escherichia coli*. Microb. Drug Resist. 2, 51–54.
[11] Bertrand, J.A., Auger, G., Martin, L., Fanchon, E., Blanot, D., Le Beller, D., van Heijenoort, J. and Dideberg, O. (1999) Determination of the MurD mechanism through crystallographic analysis of enzyme complexes. J. Mol. Biol. 289, 579–590.
[12] Anderson, M.S., Eveland, S.S., Onishi, H.R. and Pompliano, D.L. (1996) Kinetic mechanism of the *Escherichia coli* UDPMurNAc-Tripeptide D-alanyl-D-alanine-adding enzyme: Use of a glutathione *S*-Transferase fusion. Biochemistry 35, 16264–16269.
[13] Pratviel-Sosa, F., Acher, F., Trigalo, F., Blanot, D., Azerad, R. and van Heijenoort, J. (1994) Effect of various analogues of D-glutamic acid on the D-glutamate-adding enzyme from *Escherichia coli*. FEMS Microbiol. Lett. 115, 223–228.
[14] Auger, G., van Heijenoort, J. and Blanot, D. (1995) Synthesis of *N*-Acetylmuramic acid derivatives as potential inhibitors of the D-glutamic acid-adding enzyme. J. Prakt. Chem. 337, 351–357.
[15] Gegnas, L.D., Waddell, S.T., Chabin, R.M., Reddy, S. and Wong, K.K. (1998) Inhibitors of the bacterial cell wall biosynthesis enzyme MurD. Bioorg. Med. Chem. Lett. 8, 1643–1648.
[16] Zeng, B., Wong, K.K., Pompliano, D.L., Reddy, S. and Tanner,

M.E. (1998) A phosphinate inhibitor of the *meso*-diaminopimelic acid-adding enzyme (MurE) of peptidoglycan biosynthesis. J. Org. Chem. 63, 10081–10086.

[17] Auger, G., van Heijenoort, J., Vederas, J.C. and Blanot, D. (1996) Effect of analogues of diaminopimelic acid on the *meso*-diaminopimelate-adding enzyme from *Escherichia coli*. FEBS Lett. 391, 171–174.

[18] Bugg, T.D. and Walsh, C.T. (1992) Intracellular steps of bacterial cell wall peptidoglycan biosynthesis: Enzymology, antibiotics, and antibiotic resistance. Nat. Prod. Rep. 9, 199–215.

[19] Auger, G., Martin, L., Bertrand, J., Ferrari, P., Fanchon, E., Vaganay, S., Pétillot, Y., van Heijenoort, J., Blanot, D. and Dideberg, O. (1998) Large-scale preparation, purification, and crystallization of UDP-*N*-Acetylmuramoyl-L-alanine:D-glutamate ligase from *Escherichia coli*. Protein Expr. Purif. 13, 23–29.

[20] Sambrook, J., Fritch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

[21] Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E.W., Lory, S. and Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406, 959–964.

[22] El-Sherbeini, M., Geissler, W.M., Pittman, J., Yuan, X., Wong, K.K. and Pompliano, D.L. (1998) Cloning and expression of *Staphylococcus aureus* and *Streptococcus pyogenes murD* genes encoding uridine diphosphate *N*-acetylmuramoyl-L-alanine:D-glutamate ligases. Gene 210, 117–125.

[23] Wong, K.K., Kuo, D.W., Chabin, R.M., Fournier, C., Gegnas, L.D., Waddell, S.T., Marsilio, F., Leiting, B. and Pompliano, D.L. (1998) Engineering a cell-free murein biosynthetic pathway: Combinatorial enzymology in drug discovery. J. Am. Chem. Soc. 120, 13527–13528.

[24] Reddy, S.G., Waddell, S.T., Kuo, D.W., Wong, K.K. and Pompliano, D.L. (1999) Preparative enzymatic synthesis and characterization of the cytoplasmic intermediates of murein biosynthesis. J. Am. Chem. Soc. 121, 1175–1178.

[25] Wulfing, C., Lombardero, J. and Pluckthun, A. (1994) An *Echerichia coli* protein consisting of a domain homologous to FK506-binding proteins (FKBP) and a new metal binding motif. J. Biol. Chem. 269, 2895–2901.