Argumentation in biology: exploration and analysis through a gene expression use case

Kenneth C. M^{c}Leod

Doctor of Philosophy in Computer Science

Heriot-Watt University, Department of Computer Science.

June 2012

Abstract

Argumentation theory conceptualises the human practice of debating. Implemented as *computational argumentation* it enables a computer to perform a virtual debate. Using existing knowledge from research into argumentation theory, this thesis investigates the potential of computational argumentation within biology.

As a form of non-monotonic reasoning, argumentation can be used to tackle inconsistent and incomplete information - two common problems for the users of biological data. Exploration of argumentation shall be conducted by examining these issues within one biological subdomain: *in situ* gene expression information for the developmental mouse.

Due to the complex and often contradictory nature of biology, occasionally it is not apparent whether or not a particular gene is involved in the development of a particular tissue. Expert biological knowledge is recorded, and used to generate arguments relating to this matter. These arguments are presented to the user in order to help him/her decide whether or not the gene is expressed.

In order to do this, the notion of *argumentation schemes* has been borrowed from philosophy, and combined with ideas and technologies from artificial intelligence. The resulting conceptualisation is implemented and evaluated in order to understand the issues related to applying computational argumentation within biology.

Ultimately, this work concludes with a discussion of Argudas - a real world tool developed for the biological community, and based on the knowledge gained during this work.

Acknowledgements

Many people gave expert advice and support during this work. In particular the efforts of the following should be acknowledged:

Dr Jeff Christiansen - main biological expert;

Dr Shanmugasundaram Venkataraman - secondary biological expert;

Dr Ian Overton - biological expert;

Dr Yiya Yang - EMAGE database and programmatic access expert;

Staff of the Human Genetics Unit - evaluation subjects;

Dr Gus Ferguson - evaluation expert;

Prof Patrik Holt - evaluation expert;

Dr Albert Burger - PhD supervisor.

HERIOT WATT UNIVERSITY

ACADEMIC REGISTRY
Research Thesis Submission

Name*:*	Kenneth Campbell M^{c}Leod		
School/PGI:	School of Mathematical and Computer Sciences		
Version: *(i.e. First, Resubmission, Final)*	Final	Degree Sought (Award **and** Subject area)	PhD Computer Science

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	

Submission

Submitted By *(name in capitals):*	Kenneth M^{c}Leod
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
*E-thesis Submitted (**mandatory for final theses)***			
Signature:		Date:	

Please note this form should bound into the submitted thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Argumentation theory [1] is a multidisciplinary field that studies arguments and arguing.

An *argument* is a reason to believe something is true. This may be a formal proof, or a piece of natural language: for example, a reason to carry an umbrella.

Philosophers, e.g. Walton [2], who wish to study arguments in greater depth use the notion of argumentation schemes [3] to classify individual arguments. A scheme is essentially a natural language template in the form of a if-then rule: *if* X, Y, and Z are true, *then* A may be true too. Additionally, this layout enables schemes to be used for the generation of arguments.

The crucial attribute of an argument is its *defeasibility*. Defeasible is a term originating from the legal community meaning that an argument is "open in principle to revision, valid objection, forfeiture, or annulment" [4]. The word *defeasible* is derived from the Old French word *desfesant* meaning 'undoing' [4]. As such an argument may provide a reason to believe something is true, but it does not prove it is definitely true. Defeasibility is important because arguments can be in conflict: one argument can suggest a conclusion is true whilst a second argument intimates that the same conclusion is false. Often it is desirable to make a decision about the validity of a conclusion. As the evidence and corresponding arguments change over time, it may be necessary to revise that decision. Defeasibility enables this.

Arguing (commonly referred to as *argumentation* in the literature) is the process of using arguments to justify a point of view. This process may take place between multiple agents (human or software) inside a debate, or it may be carried out by a single agent: e.g. a politician justifying the government's decision to increase taxes.

The idea of argumentation is natural for humans. Debating who will win the weekend's big football match, discussing whether or not it will snow, quarrelling with a partner over whose turn it is to stack the dish washer - all of these are typical examples of argumentation. However, it is a different form that is of interest here.

Envisage two medics discussing the treatment of a patient: one may recommend a particular course of action whilst the second prefers a different approach. Ideally, each medic will put forward arguments supporting their favoured option, and possibly highlight a weakness with the other's choice. The medics will listen to their colleague's view, and express an honest critique of the arguments advanced. The debate will be fair, reasoned, and (where possible) based on solid medical research. Two important differences exist between this example and the work presented in this thesis; firstly the domain is biology not medicine, and secondly the agents will be virtual.

The subdomain of *computational argumentation* involves the use of computers for constructing and using arguments. The range of domains in which argumentation has been applied is broad, for example: artificial intelligence & law [5], ontology matching [6], decision support systems [7], and agent communication [8].

Argumentation in relation to biology is surprisingly rare. Jefferys *et al.* [9] use argumentation to analyse the output of a protein prediction tool. However, most work involves pedagogical efforts to improve the construction of natural language scientific arguments by students, e.g. Adúriz-Bravo *et. al.* [10].

Although the study of argumentation is rare in biology, it is common within the medical world. A wide range of decision support systems exist, for example Fox *et. al.* [7]. Additionally, argumentation can be used to persuade patients to change their behaviour [11], and generate pamphlets to explain complex issues to patients [12]. Although clear differences exist between medicine and biology, there are many similarities. If argumentation has been shown to work well in one field, intuitively it should work in the other.

The overarching goal of this work is to explore argumentation within biology, with the focus on a single use case - *in situ* gene expression information for the developmental mouse. *Genes* are units of biological instructions that direct the body on what to build, and how to do it. *In situ* gene expression is concerned with discovering where the genes are present in a particular organism. In this use case the organism is the unborn (so-called *developmental*) mouse. Like most biological subdomains, the information is distributed across a number of resources and is both incomplete and inconsistent.

1.2 Summary of thesis

The ambition of this work is to improve the understanding of the applicability of computational argumentation within biology through the study of a particular biological use case involving in situ *gene expression for the developmental mouse. Whilst tackling the problems faced in the use case, insights into the future role of computational argumentation in biology shall be sought.*

This will be achieved through the following activities:

Creation of argumentation schemes to record the knowledge of an expert biologist within the use case. The knowledge will be converted into logical inference rules for use in a third party argumentation toolkit.

Consideration of what *argumentation* means within the current use case. Which approach makes most sense? What constraints are shown to restrict the solutions available?

Design of an architecture suitable for argumentation systems in the biological domain. In the biological domain data is commonly spread across a number of resources, and requires integration.

Implementation of a series of prototypes that collectively enable the ideas contained in this thesis to be tested. Associated evaluations ensure that the lessons can be documented.

Exploration of argument presentation in order to improve the understanding of the manner in which biologists want arguments and argumentation to be presented.

Review of the work undertaken and lessons learnt for future argumentation activities within biology.

1.3 Contributions

Important outcomes of this work include both academic and domain community contributions; they are both summarised in the following list of bullet points before being discussed at length afterwards:

- Creation of argumentation schemes to document expert knowledge relating to the analysis of gene expression information;

- Development of an architecture for an argumentation-based platform, which aggregates distributed data before arguing with it;

- Investigation of the mechanisms with which argumentation can be presented to a biological audience. Encompassing evaluation, analysis and direct comparison of a number of presentation forms including natural language representations and graph-based visualisations;

- Exploration of a gene expression use case leading to a consideration of the application of argumentation in the biological domain including an assessment of the role of the wider notion of argumentation theory plus the application of computational argumentation;

- Review of the issues that hinder the development of argument-based solutions within biology;

- Proposal of ideas for the development of both argumentation and biology to enable further penetration of argumentation-related technology into the biological domain;

- Implementation of a real world tool that implements, evaluates, and evolves the ideas discussed in this work;

- Frank consideration of the future role of the notion of argumentation within the life sciences, including the genesis of a number of key questions regarding the future of this technology.

Expert knowledge for the domain of *in situ* gene expression information for the developmental mouse has been extracted and documented. To date, this thesis appears to be the only collection of such knowledge that is not geared towards domain experts. The application of argumentation schemes in this context provides a useful resource for those argumentation scholars who wish to investigate the philosophical aspects that this work has not explored. Additionally, the publications obtained from this process provide a pragmatic evaluation of the use of schemes in this manner. Following on from the generation of schemes, is the transformation of schemes into formal logic inference rules. Again, real world experience of applying academic theories is communicated.

During this thesis an architecture supporting argumentation with distributed data sources is developed. Analysis of this architecture supports its evolution, and enables it to be utilised as a starting point for future endeavours.

An investigation of the presentation of arguments has revealed that the biological user group has a wide variety of expectations. As such it is unlikely that any single presentation framework will satisfy everyone. Whilst some users clearly prefer written forms of argument, they do not wish to read a lot of information. Others prefer visual communication means; however, there is no single presentation form that satisfies this group. The default form of visual presentation for the argumentation community is shown to be the least suitable for a biological user group. Manifestly, this is a difficult area of research requiring a significant application of resources and expertise. This work reaches a compromise solution, which is shown to be effective for the current use case, and may provide a blueprint for future work.

The thesis provides an analysis of argumentation within the use case, and the wider biological world. Consideration is given not only to the role of computational argumentation, but to the ways in which other forms may be deployed. It is clear that there is significant potential for the application of argumentation. Yet the success of argumentation will be limited by both the nature of biological knowledge and the lack of maturity in argumentation-centric research.

Currently biological data is very fragmented and its organisation arbitrary, which creates a significant barrier to the deployment of argumentation based technologies. For many techniques, not least argumentation, their promise can never be fulfilled within this field until access to the domain's information resources is standardised. For those researchers interested in further contemplating the role of argumentation, or similar concepts, the nature of biological data is recounted in a manner that makes explicit the difficulties encountered.

Reviewing this document provides a number of ideas for future research for those interested in following behind this work. Perhaps most vital is the need for cross community teams that will develop the social and technical aspects concurrently. For example, there is clearly a requirement to consider further the presentation of both arguments, and the context in which they are used (argumentation). Equally obvious is the imperative for an improvement in the tools and methods used to help domain experts develop schemes.

Ultimately, this work generates a solution to the problems associated with the individual use case. During the lifetime of the BBSRC[1] funded Argudas project, the ideas contained in this thesis were implemented for the benefit of real users. Following a sustained dialogue with typical users, the tool evolved in a number of unexpected ways. The experience gained through the development of Argudas is documented, and drawn upon to illuminate the role of argumentation within the life sciences.

From this thesis two conclusions are immediately obvious. Firstly, argumentation is a concept with substantial potential. Secondly, the tools, methods, and ideas supporting the application of argumentation are not yet mature. Accordingly, there are a number of issues to tackle before argumentation can be easily, and successfully, used within the life sciences. These matters are considered, and where possible solutions are discussed. Where realistic resolutions are not feasible within the constraints of a PhD, the relevant topics are highlighted for future consideration.

1.4 Structure of thesis

This thesis has been organised into 13 chapters as follows.

[1] Biotechnology and BioSciences Research Council - www.bbsrc.ac.uk

Background to the field of argumentation, and a discussion of the key elements appropriate to this work, are provided in Chapter 2. This chapter attempts to portray the broad and varied constitution of the research undertaken in this field; however, it does not stray into the psychological or sociological aspects. Nor does it dwell in depth upon the philosophical or logical roots of the discipline. Instead attention is directed towards computational argumentation and the areas of its application.

Chapter 3 explores the relevant biology necessary for an understanding of this document. This includes a depiction of *in situ* gene expression techniques, the data resulting from the exploitation of those techniques, and the developmental mouse. A thorough investigation of the biological resources at the centre of the current use case provides the finish.

The problem and motivation behind this work are discussed in Chapter 4. It scrutinises the twin issues of inconsistency and incompleteness within the use case, and demonstrates that the distributed nature of the data causes frustration for biologists. Included in this review are a series of illustrative examples that clearly define the use case. This chapter finishes by defining the topics that subsequent chapters will investigate, including: the best way of presenting arguments to typical biological end users, the appropriateness of argumentation for this use case, and knowledge gleaned relevant for the application of argumentation within biology in general.

Chapter 5 presents one way of using argumentation to tackle the problems described in Chapter 4. Before this approach is tested in subsequent chapters, an overview is provided here. As part of this discussion, the use of argumentation is justified by comparing it with a number of possible alternative technologies. Concluding this chapter is a walkthrough that explores argumentation within the current use case.

The proposed solution requires the capture of expert biological knowledge, and the transformation of that knowledge into inference rules. These topics are the central themes of Chapter 6, which first refines the notion of argumentation schemes before recalling the workflow used to generate the inference rules.

To ensure the ideas contained in this thesis are both practical and applicable in the real world, two separate prototype systems are designed and implemented. The initial prototype is an instantiation of the architecture developed during this work, and acts as a basic proof of concept. The second facilitates the evaluation of the argumentation presentation mechanism developed here. Recorded within this prototype was the current understanding of the best way to present arguments, and argumentation, to a biological user. Chapter 7 discusses these systems reviewing the implemented architecture and workflow, before considering the constraints that shape the systems.

Chapter 8 documents the evaluation of the second prototype. Two separate evaluation exercises are performed. The first is a detailed inquiry into the usability of argumentation within biology, and the appropriateness of the prototype system for the expected end users. The second evaluation examines the presentation of visual arguments, endeavouring to ascertain which presentation style is most appropriate.

Chapters 9 - 11 respond to the queries raised during Chapter 4, attempting (where possible) to answer them succinctly and provide concrete recommendations for improving the uptake of argumentation within biology.

An analysis of the results from Chapter 8 is provided in Chapter 9. The emphasis is on replying to the query: what is the best way of presenting arguments to typical biological end users? Accordingly, this chapter considers the presentation of visual arguments, and whether or not textual representations should supplement visual presentation. Additionally, the appropriateness of summarising the argumentation process is reflected on.

The pertinence of argumentation for biology is explored during Chapter 10. Additionally, the applicability of the solution proposed is considered, alongside other plausible ways of exploiting argumentation. Although the analysis is use case centric, it applies equally to the wider biological world.

Reporting the knowledge gathered, especially that relevant to argumentation in the wider world of biology, is the aim of Chapter 11. The chapter primarily revolves around the capture and transformation of schemes; exploring and critiquing the workflow that provides the rules used to generate arguments. Additionally, an investigation of what is required for computational argumentation to take a stronger hold within the biological domain takes place.

An extension of the work described in previous chapters is the focus of the penultimate chapter. As part of a BBSRC funded project, the Argudas tool evolves from the solutions discussed in Chapter 7. Included is a review and justification of the changes made, and the extensions considered in order to improve performance and suitability for the biological end users. An evaluation of these changes provides the close.

The document is concluded in Chapter 13.

1.5 Published papers

The work of this thesis has been reported in the following papers.

Journal

K. M^{c}Leod, G. Ferguson, and A. Burger, Argudas: lessons for argumentation in biology based on a gene expression use case, BMC Bioinformatics, in press, 2011.

K. Sutherland, K. M^{c}Leod, G. Ferguson, and A. Burger, Knowledge-driven enhancements for task composition in bioinformatics, BMC Bioinformatics, 10(Suppl 10):S12, 2009.

K. M^{c}Leod and A. Burger. Towards the use of argumentation in bioinformatics: a gene expression case study. Bioinformatics, 24:i304i312, 2008.

Conference/Workshop

K. M^{c}Leod, G. Ferguson, and A. Burger, Argudas: arguing with gene expression information, in Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences, 10 December 2010, Berlin, Germany.

K. M^{c}Leod, G. Ferguson, and A. Burger, Using Argumentation to resolve conflict in Biological Databases, in Proceedings of the IJCAI-09 workshop on Computational Models of Natural Argument, 13 July 2009, Pasadena, CA.

K. Sutherland, K. M^{c}Leod, and A. Burger. Semantically linking web pages to web services in bioinformatics. In 3rd International Application of Semantic Technologies

Workshop, Munich, Germany, September 2008.

K. M^{c}Leod and A. Burger. Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In Proceedings of IADIS International Conference Applied Computing, pages 489492, Salamanca, Spain, February 2007. IADIS Press.

Technical report

G. Ferguson, K. M^{c}Leod, K. Sutherland, and A. Burger. Sealife evaluation. Technical Report 0063, Dept. of Computer Science, Heriot-Watt University, 2009.

Chapter 2

Argumentation

Argumentation theory - the study of arguments and arguing - is a multi-disciplinary area with strong philosophical roots dating back to the work of Aristotle, e.g. [13]. Today argumentation theory covers a wide range of fields exploring the nature and science of communication and persuasion. Work from the traditional sources - philosophy and logic - is supplemented by research in psychology, sociology and communication studies.

Artificial intelligence (AI) is central to the modern investigation of argumentation theory. It provides the mechanism to implement many of the theories produced, and an increasing range of tools that further the understanding of the domain.

This thesis takes a pragmatic approach to investigating the potential of argumentation theory for bioinformatics. Consequently, the approach will be AI-centric - the other main contributors to modern argumentation theory are less relevant, and shall not be explored in depth.

In this chapter the core of argumentation shall be explored, starting with an introduction to argumentation theory (Section 2.1), and progressing to concentrate on argumentation in AI (Section 2.2). Section 2.3 briefly describes different mechanisms for performing argumentation in AI, while Section 2.4 focuses on one - argumentation frameworks. The visualisation of arguments and argumentation is considered in Section 2.5 before applications are discussed in Section 2.6.

2.1 Argumentation theory

Although *argumentation* is used in this work to refer to the process of arguing, it is often used to refer to argumentation theory. A subject that has evolved considerably over the years:

> Argumentation, as a growing interdisciplinary field of research, was conducted mainly in logic, philosophy, and communication studies in the beginning. It has now branched and become truly interdisciplinary as it has affected more and more fields, like cognitive science, where models of rational thinking are an essential part of the research program. ([14] page 287)

Likewise there is a suitably extensive view of argumentation e.g.:

> . . . the concept “argumentation” understood as an intrapersonal, interpersonal, inter-group and intercultural process of mutual influence and decision-making ([15] page 18)

This quote highlights the role of social sciences within argumentation, and the importance of real world argumentation. Although the focus of this thesis is on the use of computers to perform argumentation, it is acknowledged that this is fundamentally a human process:

> Professionals routinely undertake argumentation as an integral part of their work. Consider diverse types of professional, such as clinicians, scientists, lawyers, journalists, and managers, who have to identify pros and cons for analysing situations prior to presenting some information to an audience and/or prior to making some decision. ([16] page 1)

To emphasise further the human nature of arguing this section continues with an informal discussion of real world argumentation.

2.1.1 Real world argumentation

Monological and dialogical are two categories of argumentation in the real world.

Dialogical argumentation is based on the notion of a dialogue - a goal directed communication in which, at least, two parties participate sequentially by responding to the previous communication from the other party [14]. Arguments supporting their alternative views on a subject are traded with the goal of changing the opinion of the other party (or audience). The ability of the second party to respond creates the dialogical aspect. Common examples include class room debates, legal procedures and medical case conferences.

Monological argumentation involves one agent. Examples include newspaper editorials, scientific reports, and political speeches. One person creates all the arguments, often presenting the case for and against the issue being discussed before deciding on a viewpoint to recommend to the audience. The audience is then able to evaluate the arguments and determine if they accept the recommended view. Unlike dialogical argumentation, there is no opportunity for a direct response or debate.

Besnard and Hunter describe the difference between the two forms as follows:

> ...in monological argumentation, the emphasis is on the final result, whereas in dialogical argumentation, the emphasis is the process as represented in terms of dialogue exchanges. ([16] page 11)

Common to both styles are the notions of *audience* and *persuasion.* In each form of argumentation the aim is to persuade the audience that what the speaker says is correct. Perelman and Olbrechts-Tyteca [17] discuss the role of the audience, highlighting that each member of the audience will bring their own views and values to bear in their evaluation of arguments, and thus one argument may persuade some members but not others.

The notion of *burden of proof* [18] is associated with dialogical argumentation. It determines which participant in a dialogue has something to prove. If the speaker makes a claim, then (s)he has a responsibility to back it up if required to do so. Likewise, a listener who rejects a claim should be able to support his/her objection.

Argumentation is only one form of communication. A single dialogue may contain many different categories - dialogues often being classified according to the system of Walton and Krabbe [19].

Research

A number of research domains focus on real world arguments, one such is *informal logic* (IL), also known as *critical argumentation* - those with a logical background favouring the former title and philosophers the latter. This subject can be defined as:

> The three goals of critical argumentation are to identify, analyze, and evaluate arguments. The term“argument” is used in a special sense, referring to the giving of reasons to support or criticize a claim that is questionable, or open to doubt. ([14] page 1)

Traditionally this field studies monological arguments, captured in a written form, preferring to leave dialogical argumentation to those interested in dialogue logic [20]. As such, it uses the *argument as product* notion. Increasingly researchers are moving towards the idea of *argument as process* as a mechanism for evaluating arguments, e.g. Walton [21].

The distinction between argument as a product and a process is based on the work of O'Keefe [22] and Habermas [23]. The former category is the traditional logical notion of an argument as a series of statements that lend some support to a particular conclusion. An argument is a unit of reasoning. Habermas called this *argument*, and O'Keefe *argument1*. Argument as process is a debate or discussion, with the purpose of reaching a conclusion about some particular claim that is initially in doubt. Now argument is a reasoning process. This is *argument2* in O'Keefe's terminology and *argumentation* in Habermas'. In this thesis Habermas' terminology is adopted.

Pragma-dialectics [24] is one of the most famous disciplines devoted to studying argument as process. It views argumentation as a:

> complex speech act which is both constituted and regulated by pragmatic rules. ([14] page 288)

Such work, together with argumentation theory in general, has shaped the view of argumentation in AI.

2.2 Argumentation in Artificial Intelligence

Traditionally, mathematical based disciplines assume that the domains they are reasoning in are consistent, complete, and monotonic. The term *consistent* indicates that none of the information in the domain is contradictory. *Complete* describes a situation where all information is known (or the necessary information/knowledge can be derived from the known information). A result of the consistent and complete nature of these domains is that they are *monotonic* - the addition of new information never causes previously inferred knowledge to be withdrawn.

The assumptions of consistency, completeness, and monotonicity are clearly flawed when considered in terms of real life. Planning, data management and decision making all need to consider situations that are incomplete, inconsistent and non-monotonic - such challenges are studied by the AI community under the heading of *non-monontonic reasoning* (NMR).

AI has proposed a variety of solutions for NMR; Reed and Grasso [25] claim that the majority of these solutions are descended from four proposals. These techniques focus on what it is possible to believe in the absence of information. For example, *default logic* [26] proposes that a generalisation is used to infer what would normally happen, e.g. birds normally fly, and Tweety is a bird, so Tweety can fly. *Nonmonotonic logic* [27] uses a different approach to perform essentially the same intuition. *Negation as failure* [28], is a mechanism to implement the *closed world assumption*: when something is not known to be true, assume it is false. In *circumscription* [29] a list of abnormal situations is described, with all other situations assumed to be normal. Then a proffered inference is taken to be true unless the situation is abnormal.

These approaches are all linked through the work of Lin and Shoham [30], which is the ancestor of Dung's seminal work [31]. Dung himself said:

> ... many of the major approaches to nonmonotonic reasoning in AI and logic programming are different forms of argumentation. This result should not be very surprising ... all forms of reasoning with incomplete information rest on the simple intuitive idea that a defeasible statement can be believed only in the absence of any evidence to the contrary which is very much like the principle of argumentation. ([31] page 4)

Furthering the idea of tying together the various forms of NMR, the *assumption-based framework* [32] provides a way to model all major implementations of NMR including argumentation.

Argumentation is not limited to merely NMR, it has emerged as a general computational paradigm. The point of crossover is often attributed to the work of Dung [31]:

> ...the model of argumentation described in [31] is now recognised as providing an important bridge between argumentation theory as a supporting analytic tool for non-monotonic reasoning and the independent exploitation of argumentation models in wider AI contexts. ([1] page 4 - reference style changed to that of this dissertation[1].)

Reed and Grasso [25] chart the history of argumentation in AI from the early days in which it considered NMR, through the work of Dung, and on towards the use of argumentation in an expanding range of AI problems. They split systems into two broad categories: those that *model with argument*, and those that perform *modelling of argumentation*. The former category includes work where the concepts of arguing and arguments are applied to situations that are not normally associated with them, for example the work on NMR.

Modelling of argumentation is the idea of modelling natural argumentation - that is human based argumentation. Much of this work has taken place in the field of AI and law (see Section 2.6.1). Additionally it includes: visualisation tools that help diagram arguments, e.g. Reed *et. al.* [33]; tools for collaborative work, e.g. de Moor and Aakhus [34]; decision making systems in a variety of domains, e.g. Glasspool *et. al.* [35]; and systems for educational purposes, such as giving dietary advice, e.g. Grasso *et. al.* [11].

Work on multi-agent systems (e.g. Rahwan *et. al.* [8]) and the prevalence of dialogue games in many applied areas of argumentation (e.g. AI and law) has increased the importance of the dialogue and the dialectical aspects of argumentation, prompting the field of *computational dialectics* [36].

Some of the above systems encompass a wide definition of argumentation that

[1] In the future, the style of reference will be changed without comment.

includes all aspects of argumentation theory in order to create and present arguments in a natural manner to the user, e.g. Grasso *et. al.*[11]. Others have a more restricted definition, similar to:

> Argumentation is the process of putting forth arguments to determine the acceptability of propositions. ([37] page 1)

Still, it may be more accurate to suggest that:

> . . . argumentation systems are not concerned with truth of propositions, but with justification of accepting a proposition as true. ([38] page 16)

Increasingly though:

> Argumentation processes are typically dialogs involving two or more agents, but may also be monologs. Procedural norms, called argumentation protocols, regulate the process, to help promote values such as rationality, fairness, efficiency, and transparency. ([37] page 1)

Loui [39] goes further, suggesting that argumentation can be rational only if it is conducted inside a fair and effective dialectic disputation protocol. This takes argumentation in AI closer to real world argumentation, e.g. a legal debate in which the procedural rules are written in law and enforced by a judge.

From a computational viewpoint, argumentation is normally perceived as the creation and evaluation of arguments, to determine which stance (for/against a particular claim) has the strongest arguments. That position is taken to be true. Often such a system is wrapped inside dialectical procedural models, e.g. dialogue games, in order to create a more natural process.

2.3 Methods of arguing

Removing the procedural wrapping, and considering solely the creation and evaluation of arguments, Besnard and Hunter [16] identify three basic techniques through which argumentation is performed. Their preferred approach, called a *coherence system*, removes conflict and reasons with consistent subsets of information. Using a logic that allows the inference and later retraction of a conclusion, a so-called *defeasible* *logic* such as DefLog [40], is the second mechanism featured. The third category discussed is *abstract systems*, the most famous being the seminal work of Dung [31]. Others, e.g. Prakken [41], would call these *argumentation frameworks*.

Such argumentation frameworks represent the most commonly applied mechanism for conducting argumentation in AI. Consequently, they shall be explored in more detail in the following section.

2.4 Argumentation frameworks

As noted by Prakken [41], the definition of the term *argumentation framework* is not always consistent:

> ...the term 'framework' will be used to denote the general model, to highlight that it can be instantiated in various ways (such instantiations will in turn be called argumentation systems). This contrasts with Dung's [31] use of the term 'argumentation framework', which denotes a specific set of arguments with a specific attack relation. (page 95)

This work shall adopt the terminology proposed by Prakken.

Essentially, argumentation frameworks provide an abstract specification of a system that allows computers to conduct argumentation automatically. Different styles of framework exist; Prakken [42] identified five common traits: an underlying logical language; the concept of an argument; the idea of conflict between arguments; a definition of defeat between arguments; and a method for determining if an argument is justified. Each layer builds upon the previous one, as summarised in Figure 2.1.

Layers one and two are common to most standard logical systems, it is the subsequent three layers that differentiate argumentation frameworks.

Not all frameworks specify every feature, the most obvious example being Dung [31], which only defines the final layer. There is no need to describe everything because:

> Argumentation frameworks [31, 43] for aggregating and evaluating arguments depend only on relationships between arguments, not the method used to construct them. ([37] page 2)

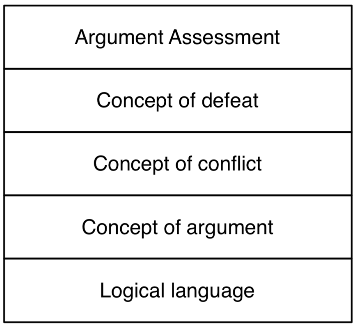


Figure 2.1: Prakken's argumentation framework from [42].

Frameworks, like Dung's, are often said to represent *abstract argumentation*. Other authors, for example Prakken [41], specify the structure of an argument (and thus all subsequent levels) - work of this nature is referred to as *structured argumentation*.

The following subsections run through the main layers of an argumentation framework.

2.4.1 Logic

Prakken's description of an argumentation framework begins with the specification of a logical language on which the rest of the system is developed. Though any logic can be used, Prakken recommends the use of a monotonic logic [42]: because defeated arguments can be reinstated they should be retained.

Often the logics utilised have a defeasible inference, e.g. Verheij's DefLog [40]. In this way two of the three methods of performing argumentation discussed in Section 2.3 can be linked. It is possible to use more powerful logics with strict (i.e. non-defeasible) implication such as classical logic [44], e.g. Pollock [45]. Unfortunately, this makes it harder to capture the defeasible nature of arguments:

> ...In classical logic, any inference can follow from an inconsistent set of assumptions. ...Thus, inconsistency causes classical logic to collapse. ([16] page 16 - 17)

2.4.2 Arguments

Central to argumentation theory is the notion of an *argument*. This term is overloaded, with many domains having specialised meanings - in computer science, an argument is the name given to a piece of input passed into a program or sub-program. These definitions shall be ignored in this thesis.

Even within argumentation theory the term *argument* has a very broad range of definitions, and thus usages, as illustrated by:

> There are, as might be expected, almost as many definitions of argument as there are argument theorists. At one end, the all-encompassing taxonomy of Gilbert [46] covers a panoply of situated action that can count as argument, from artistic creation, through non-lingustic communication, to physical activity. At the other end, van Eemeren and Grootendorst's pragma-dialectics [24] associates argument with the notion of critical discussion, a closely bounded, tightly specified linguistic activity whose definition rests upon speech act theory. ([47] page 4)

As there is no canonical meaning, in this work the definition of argument will be *a reason to believe something is true.* Examples include a reason to believe a gene is expressed in a tissue, or a mathematical proof of a formula.

Often in disciplines with philosophical roots, such as AI, arguments are composed from propositions. *Propositions* are statements that can be evaluated to be true or false, e.g. *it is raining.* One proposition is taken to be the *claim* (or conclusion) of the argument, i.e. the statement that the argument supports. The remaining propositions are the *premises* (see Figure 2.2). When all the premises are evaluated to be true, the conclusion is likely to be true. The level of likelihood depends on the class of argument, generally speaking three exist: deductive, inductive, and a third category, which shall be called *defeasible* in this document.

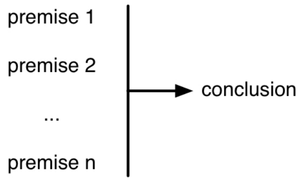


Figure 2.2: The basic layout of a logical argument.

A simple example of an argument could provide a reason why someone will get wet today:

> The weather report said it would rain today, therefore you will get wet.

This argument clearly has a premise (*The weather report said it would rain today*), and a conclusion (*you will get wet*). However, this argument does not provide a link between the premises and the conclusion, i.e. a statement of why the premises imply the conclusion. Often in natural language arguments the link is implicit, in this case:

> If it rains, you will get wet.
> The weather report said it will rain,
> therefore you will get wet.

Arguments of this style correspond to the logical inference rule known as *modus ponens*, which has the basic form: *If A then B*. Analysis of arguments, and classes of arguments, reveal that this is the pattern of many, perhaps even most, common arguments [48]. Many real world arguments need to be rewritten in order for that arrangement to become obvious - as demonstrated above.

Modus ponens style rules are widely applied in AI, where they are used for *forward chaining* [49]. Through this mechanism of chaining an individual argument can be used as a part of a bigger argument, see Figure 2.3. Arguments are linked by using the conclusion of one argument as a premise in the second. This can continue indefinitely. However, if any sub-argument is shown to be invalid then the whole argument is invalid.

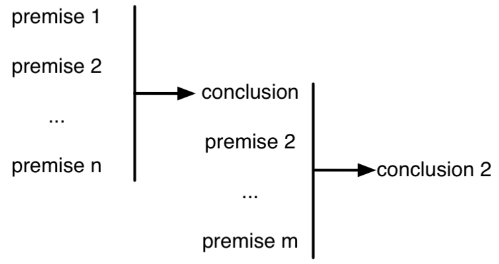


Figure 2.3: Arguments can be chained.

Argument strength

Earlier in this section, three classes of argument strength are identified: deductive, inductive and defeasible. These classes represent arguments of varying strengths, from the definitely true to the plausibly true. The stronger the argument, the greater the confidence one can have in its conclusion.

Deductive arguments are essentially proofs. They show something is true, i.e. when the premises are true the conclusion must be true. *Inductive* arguments are based on probability, and thus the premises being true implies the conclusion is probably true. Inductive arguments are weaker than deductive ones as there is a possibility that the premises are true whilst the conclusion is false. The third category is more difficult to classify, one definition is:

> forms of reasoning that are often necessary, but are more tentative in nature and need to be judged circumspectly by reserving some doubts. Such reasoning is presumptive and defeasible. This kind of reasoning is only plausible and is often resorted to in conditions of uncertainty and lack of knowledge. ([2] page 10)

An argument is plausible when it appears to be true [50], i.e. plausible inferences seem reasonable, but cannot be verified. Such arguments are the weakest of the three classes.

According to Walton:

> A presumption is something that can be accepted by agreement temporarily as things go forward unless at some future point in the exchange it is shown to be unacceptable. ([50] page 166)

For example: "innocent until proven guilty". From the definition, it seems that a presumption is based on a plausibility that in the future may be proven false, thus overturning the presumption.

The term *defeasible* was introduced to the philosophical world by Hart [51] through his work in law. Traditionally it refers to arguments that are rationally compelling but not deductive. Today its meaning is more restricted:

> In recent work, the term defeasible reasoning has typically been limited to inferences involving rough-and-ready, exception-permitting generalizations, that is, inferring what has or will happen on the basis of what normally happens. [52]

This definition suggests that something can be presumed true "on the basis of what normally happens" unless there is an unusual circumstance (i.e. *exception*) indicating that this is not a normal situation.

There appears to be an overlap between the definitions of presumption and defeasibility, with both intertwined with the notion of plausibility. A clarification of these terms is beyond the scope of this document. Instead, the third group of arguments, lacking the verifiable backing of probability and thus the strength of inductive arguments, shall be called *defeasible* to emphasise their fragile nature.

Argument acceptability

Earlier, to define argumentation, the following quote was used:

> Argumentation is the process of putting forth arguments to determine the acceptability of propositions. ([37] page 1)

According to this quote, the ability to determine if an argument is true or false is central to argumentation. Details of the mechanisms for performing this classification are reserved until Section 2.4.6; however, the abstract notion shall be discussed further in this section.

Acceptable arguments are those that the evaluator (software or human) finds to be correct (*true*) - as such these are the arguments with which a dialogue can be won. Pollock [53] defines two types of acceptability. Firstly, an argument is *justified* if it appears to be acceptable at the current moment. It becomes *warranted* if the argument would be acceptable if the agent evaluating it had infinite resources to apply. This distinction is not universally employed, with most authors seeming to concentrate on what Pollock calls justified arguments. Terms used to indicate acceptance include *valid*, *undefeated*, *justified*, *in force*, and, *preferred* amongst others.

If there are acceptable arguments, intuitively there are some that are unacceptable, i.e. believed to be false. *Invalid*, *defeated*, *overruled*, *not in force*, and *not preferred* are some of the terms used to describe this group.

A third group, sometimes called *defensible* arguments [38], may be added. These are arguments that are neither valid nor invalid.

2.4.3 Classes of argument - argumentation schemes

This subsection breaks away temporarily from the overall theme of argumention frameworks in order to discuss argument hierarchies and how they can be applied in frameworks.

In the same way that object oriented programming [54] uses the notion of a *class* to define standards for a set of similar concepts, the notion of an *argumentation scheme* defines standard arguments. This thesis employs the notion of schemes created by Walton that has evolved from [55] to [2] in response to further investigation by Walton, and in light of criticism from peers such as Blair [56].

The history of argumentation schemes is long, with Grassen [3] tracing the lineage of modern work by Walton and others back through the writings of Kleinpointner [57], Hastings [58], Perelman and Olbrechts-Tyreca [17], and more, to eventually reach Aristotle's work on Topics [13]. Many of these authors present their own notion of a scheme and classification system for it. Today, there is no agreed formalism or classification system for schemes. Most research describes schemes in a manner similar to the following:

> Argumentation schemes are forms of argument (structures of inference) that represent structures of common types of argument used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation. They include the deductive and inductive

> forms of argument that we are already so familiar with in logic. However, they also represent forms of argument that are neither deductive nor inductive, but that fall into a third category, sometimes called defeasible, presumptive, or abductive. ([2] page 1-2)

Following the work of Walton [55] schemes are associated mostly with defeasible arguments. Schemes are now widely used in fields such as communication studies, IL and AI.

An example scheme, from Walton [55], is presented in Figure 2.4. It captures the notion that a user can trust the opinion or conclusion of someone who probably knows the answer to a query (s)he has. For example, if you are lost in Tokyo you may ask a local person where you are as they probably know the answer. This argument is defeasible because there is no guarantee that the person knows where you are, or that they will answer truthfully. In Figure 2.4 the letters *a* and *A* represent variables that can be instantiated, doing so generates an instance of the scheme, i.e. an argument.

a is in a position to know whether *A* is true (false).

a asserts that *A* is true (false).

Therefore *A* is true (false).

Figure 2.4: An argumentation scheme: argument from position to know [55].

Because of Hastings [58], each scheme is associated commonly with a range of *critical questions* (Figure 2.5 provides the questions for the scheme in Figure 2.4). The exact role and function of the critical questions is still open to philosophical debate [59]. Many consider the questions to be a heuristic or memory tool guiding the evaluation of an argument, e.g. Walton [55]. In such a role the questions help a user verify that an argument is both convincing and a valid instance of the scheme. However, the application of the documented critical questions is rarely sufficient to prove the acceptability of an argument.

The uses of schemes

Broadly speaking, schemes can be used for the classification and evaluation of existing arguments, or the generation of new arguments.

1. Is a in a position to know whether A is true (false)?

2. Is a an honest (trustworthy, reliable) source?

3. Did a assert that A is true (false)?

Figure 2.5: Argument from position to know: critical questions [55].

Classifying an argument involves determining which scheme it is an instance of. An argument must be dissected to find the premises and conclusion. These propositions may need to be simplified and rewritten in order to fit a scheme. The content and purpose of the argument must be compared to those of the schemes, and the critical questions examined to ensure they are appropriate. Such is the variety and complexity of schemes that this may be a difficult task. Once classified it is possible to use the scheme to reconstruct and/or evaluate the argument.

Reconstruction is common because many written and spoken arguments are incomplete with a premise or conclusion left unsaid - in logic these are called *enthymemes* [60]. If the argument has been classified it is possible to determine what the missing statement should be, and thus reconstruct the full argument. However, doing so may misrepresent what the arguer meant and should be considered potentially misleading [61].

Evaluation of an argument is undertaken to determine if the argument is valid (*cogent*). The role of a scheme, and its critical questions, in the evaluation of an argument is summarised as:

> The scheme identifies the form of the argument and its premises. Once the premises have been identified, they can then be questioned to see if there is support for them. The critical questions indicate other ways in which the arguments can be questioned or criticised by indicating key assumptions that the worth of the argument depends on. The asking of a critical question throws doubt on the structural link between the premises and the conclusion. ([2] page 15)

Philosophically rooted debate continues into the exact role and nature of the critical questions, e.g. [20, 59]. This will impact on the implementations produced in AI. However, the above approach is not the sole mechanism for evaluating arguments in AI.

Tangentially, within AI, schemes are being used increasingly to create arguments: the schemes represent templates that produce arguments when their variables are instantiated. For example, Grasso *et. al.* [11] use Perelman and Olbrechts-Tytecas' interpretation of schemes [17], to create arguments that persuade a human to eat more healthily.

Implementation of argumentation schemes

To employ argumentation schemes with a system, both in general and within this thesis, the work of Verheij is utilised.

Verheij [48] approaches schemes through his work in AI & law. He suggests that most schemes appear to have a form similar to *modus ponens* in which a series of premises imply a conclusion is true. Further, as the arguments he examines are defeasible and *modus ponens* is not, he proposes a defeasible form called *modus non expliciens*. Many authors, such as Walton [60], rename this defeasible *modus ponens*. Verheij [62] shows this new form of *modus ponens* makes it possible to model schemes as defeasible inference rules - Verheij does this in his own language DefLog [40].

Subsequently, Verheij models critical questions in logic. His analysis of the usage of critical questions in the literature, [62], leads him to believe they fulfil four functions: restating the premises of the scheme; adding extra conditions; highlighting exceptions to the use of the scheme; and finally, suggesting other arguments that may be relevant. He proposes that the first role can be ignored as any formal system should ensure the premises are true. The second role essentially adds extra conditions, or premises, to the defeasible inference rule. The exceptions can be modelled as a Pollock undercutter (see Section 2.4.5 for an explanation), with the other arguments being modelled in new schemes and thus new rules. A similar approach is given in Bex *et. al.* [63].

Schemes can be used in argumentation systems by converting the schemes into defeasible inference rules. Verheij's method allows many existing argumentation frameworks to use schemes, even though they are not explicitly catered for, because these frameworks sit upon a logical language that permits the use of defeasible inference rules.

This is perhaps the most common way of using schemes in a framework, although others exist. For example, Gordon [37] views them as computational models for argument generation producing a search space that can be explored heuristically for acceptable arguments that support (or defeat) a particular conclusion. The evaluation of arguments is carried out via his *Carneades Argumentation System* [43].

Schemes may be used in other argumentation theory tools. An increasingly popular XML based representation of arguments, *Argument Interchange Format* (AIF) [64], can be used to model them [65]. This allows arguments and schemes to be shared between tools, for example a framework plus an argument visualisation tool like *Araucaria* [66].

2.4.4 Attack

Section 2.4.2 presents two views of arguments. Initially, a real world view in which arguments are fashioned from natural language. Secondly, a logical view in which arguments are described as a series of propositions that form premises and a conclusion. The two separate views are linked through Verheij's work [62], which shows that natural arguments can be represented in frameworks as an instance of an inference rule.

The arguments in argumentation frameworks are traditionally of a logical form (although some model arguments in other ways, e.g. Gordon has modelled them as a functional specification of a computer program [67]). Logical forms can be created by classifying real world arguments as schemes and performing the conversion, or they may be created by directly modelling the arguments as a series of logical statements. If the assumption is that the arguments are now in a standard logical form - strict or defeasible *modus ponens* - there is a finite number of ways in which they can be attacked. The terms *attack*, *conflict*, and *counterargument* are all related to the principle of one argument directly attacking another.

Figure 2.6 repeats the earlier diagram shown in Figure 2.2. If this argument is strict *modus ponens*, i.e. deductive, the argument may be attacked by contradicting the premises. If it is defeasible it is possible to attack the premises, the inference linking the premises to the conclusion, and the conclusion.

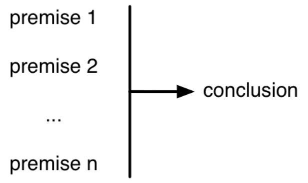


Figure 2.6: The basic layout of a logical argument.

The two main forms of attack are *rebuttal* and *undercut*. If one argument claims a gene is expressed in a tissue, and a second claims it is not, these arguments are said to *rebut* one another. Rebuttal is a direct opposition to the conclusion or a premise[2] of a second argument.

The philosopher Pollock, e.g. [68], is closely associated with the notion of *undercut*. He describes an undercutter as follows:

> For instance, suppose x looks red to me, but I know that x is illuminated by red lights and red lights can make objects look red when they are not. Knowing this defeats the defeasible reason, but it is not a reason for thinking that x is *not* red. After all, red objects look red in red light too. This is an *undercutting defeater*. Undercutting defeaters attack the *connection* between the reason and the conclusion rather than attacking the conclusion directly. ([69] page 3)

As it attacks the inference rule, Pollock's undercut can attack only defeasible arguments. One important difference between rebuttal and undercut is that rebuttal is symmetric - if argument A rebuts argument B, then B also rebuts A. Yet, if A undercuts B, B does not automatically respond.

Regrettably, Pollock's is not the only definition of an undercut, e.g.:

> Each undercut is a counterargument that contradicts the premises of an argument. ([16] page *viii* - x)

The terminology is further complicated as Pollock's notion of undercut corresponds to Toulmin's notion of rebuttal [70]. In this thesis, Pollock's terminology is favoured,

[2] Some authors, e.g. Prakken [41], prefer to classify this as third form of attack. Prakken names it *undermining*.

with no distinction made between rebutting a premise and rebutting a conclusion - both are thought of as rebuttals.

Prakken and Vreeswijk [38] identify a third form of attack called *assumption attack*. This applies to logics that include the notion of an *assumption* - there is no mechanism to prove that X is true (or false) so X is taken to be false (or true). Assumptions are found in some logics such as Default Logic [26] and have been used in argumentation frameworks, e.g. Prakken and Sartor [71]. To attack an assumption, it is necessary to show it is wrong - if X is assumed to be true, one must prove it is false. With the assumption proven incorrect, any inference based on that assumption fails. This notion of an assumption corresponds to weak negation.

Furthermore, Prakken and Sartor [38], define two methods of attack: direct; and indirect. *Indirect* attacks are against sub-arguments, with the idea that if a sub-argument is defeated, the total argument is defeated. *Direct* attacks are launched against the final argument.

Most systems for argumentation consider attack as the only relationship between two arguments, although *bipolar frameworks* [72] add the concept of *support*. Bipolarity is useful in decision making where it can be used to provide arguments that both support and oppose a decision, e.g. Amgoud and Prade [73].

2.4.5 Defeat

In order to analyse the validity of an argument it is necessary to first resolve the conflicts in which the argument is involved. Settling a conflict involves determining which argument is most persuasive. If an argument is "better" than an opposing argument it is said to be *stronger*. In real world argumentation the audience measure the strength of arguments and decide which is strongest.

Within an argumentation framework it is the fourth layer that performs this task. It examines all conflicts in the argumentation process and, for each conflict, attempts to decide which argument wins. Resulting in a series of binary relationships each stating that argument X is in conflict with, and stronger than (or at least as strong as), argument Y:

> The notion of defeat is a binary relation on the set of arguments. It

> is important to note that this relation does not yet tell us with what arguments a dispute can be won; it only tells us something about the relative strength of two individual conflicting arguments. The ultimate status of an argument depends on the interaction between all available arguments ([38] page 13)

This subsection discusses defeat in terms of individual conflicts, and not the overall status of the arguments. It is the the fifth layer of the framework that awards a status to individual arguments. It does this by creating a network (argument graph) through which it is able to determine which arguments are defeated, which are reinstated, and which are undefeated. More details on this can be found in Section 2.4.6.

Returning to the notion of defeat between pairs of arguments, Prakken and Vreeswijk [38] state that there are two basic forms of defeat: weak and strict. *Weak defeat* occurs when the attacker is not weaker than the argument being attacked. Where the arguments are the same strength they defeat each other thus cancelling each other out. *Strict defeat* describes the situation where the attacking argument is stronger than the attacked argument - the attacked argument is defeated, and the attacker is unblemished. For instance, if argument A is weaker than argument B it cannot rebut or undercut it successfully; however, B can undercut and/or rebut A. Some frameworks, e.g. Prakken and Sartor [71], deem that assumption attacks and undercuts always work, thus only rebuttals require the individual argument strengths to be compared. Prakken and Sartor justify this on the basis that it is the reality in their chosen application domain, law.

A feature of Prakken and Sartor [71] is that they use both strict and defeasible inferences. A strict inference can produce a strict argument that cannot be defeated. However, any argument that contains a defeasible inference, or piece of defeasible information, is automatically defeasible even if the rest of the argument is strict.

Conflict resolution

Generally speaking there are two paradigms for conflict resolution. In the first the conflicting arguments are compared in isolation. Arguments are balanced with all related (supporting and attacking) arguments in the second approach.

As an example of the latter technique Pollock [69] documents the idea that several weak arguments can combine to defeat a single strong argument, so-called *collaborative defeat*. In an extension of this idea, an argument that is weaker than an opponent can still diminish that opponent. Both of these approaches rely on the arguments having an associated strength.

Additionally, the concept of strength is utilised when arguments are compared directly. An argument can defeat another if it is at least as strong as it (*weak defeat*); being stronger than the other argument results in *strong defeat*.

In Pollock [69], each component of an argument (premise, inference rule, sub-argument) has a *weight* assigned to it, indicating the degree of confidence in that component. Subsequently, the *weakest link principle* is used to aggregate these weights to produce a single score for the global argument. This principle captures the idea that an argument is as strong as its weakest component, and thus the lowest weight from the components is assigned to the global argument. An alternative mechanism is the *last link principle* - the argument's strength is equivalent to the strength of the final inference rule used in the argument's creation. Krause *et. al.* [74] propose that the weight of an individual argument could be a probability, and thus the strength of the global argument computed using probability calculus.

Another mechanism to compare arguments in isolation is the notion of *preferences*. Instead of computing the individual strengths of arguments, this mechanism decides that one argument is preferable to another. This is done on the basis of a list of binary relations indicating a preference between arguments, often on the basis of a preference between premises and/or inference rules in the arguments. Such systems, called *preference-based argumentation frameworks* [75], are an extension of Dung's framework and only allow one argument to strictly defeat another if it is preferred over it. Prakken and Sartor [71] take the notion of preferences a step further by allowing them to be the subject of argumentation.

A popular mechanism for calculating such preferences is *specificity* - the most specific argument wins any conflict [76]. Simari and Loui [76] calculate their preference order using implicit information from the knowledge base. It is possible that the original data source has an explicit degree of confidence associated with the data, and this could be used in the creation of a preference ordering.

Many argue that the mechanism for deciding between arguments should be domain specific [77, 71]. For example, only a legal application will apply principles such as *lex posterior* which captures the notion that newer laws override earlier ones.

An alternative to preference-based frameworks is *value-based argumentation frameworks* [78]. These extend Dung's framework by taking account of Perelman and Olbrechts-Tyteca's notion of audience [17]. Arguments are evaluated based on the values and beliefs of individual audience members. Value-based systems ensure attacks only succeed if the value promoted by the attacker is preferred to the value promoted by the attacked; however, the success is relative to each individual in the audience as each person is likely to have their own preference ordering for the values. For example, a medic will believe effective treatment is more important than cost-effective treatment, and thus accept arguments that promote the notion of effective treatment when they are attacked by arguments that promote a cheaper, less effective alternative.

2.4.6 Evaluation of arguments

In this document, and the fifth layer of an argumentation framework, *evaluation* is treated purely as the process of determining if an argument is acceptable. Additionally, in argumentation theory it may refer to the tasks of classification, and reconstruction.

Different branches of argumentation theory evaluate arguments in various ways as they have differing purposes. In this subsection the focus is on the approach of Dung traditionally favoured in AI.

A defeated argument is one that is not accepted (i.e. regarded as false or *invalid*). In IL an argument that appears to be valid, but actually is not, is traditionally called a *fallacy* - the precise definition of this term is still being explored [79].

Broadly speaking, argumentation theory takes the suggestion of Johnson and Blair [80] that a valid argument is one where: the premises are rationally acceptable; they are relevant to the conclusion; and, they are sufficient for the conclusion to be accepted. Regrettably, this definition is problematic when considering defeasible arguments: because such arguments can be defeated, the premises cannot provide a sufficient reason to accept the conclusion. One possible solution, presented by Walton, involves weakening the meaning of sufficiency:

> Sufficiency doesn't mean that the respondent has to accept the conclusion of the argument, period. It means only that he needs to respond to it in an appropriate manner, which could include expressing doubts about it ([2] page 36)

Acceptability in AI

Although there are many methods for determining the acceptability of an argument, only one is considered in this text. For a discussion of some of the other ways see Prakken and Vreeswijk [38].

In 1995 Dung produced a seminal paper [31] that provided what Fox *et. al.* call "a calculus of opposition" ([7] page 38). Dung's work is purely theoretical; yet, it has subsequently been widely implemented, studied, and extended. For example, Prakken and Sartor [71] use it in combination with Pollock's work (e.g. [68]) to create a system for arguing in law, and Williams and Williamson [81] use it as the basis for generating explanations to be communicated to cancer patients.

Dung concentrates on the final layer of an argumentation framework - evaluation. As input to his system, Dung assumes a set of arguments and a binary set of relations for these arguments. The binary relations indicate that the second argument is defeated by the first. The input may be modelled as a directed graph (e.g. see Figure 2.7) with the nodes being the arguments and the edges the defeat relations between them.

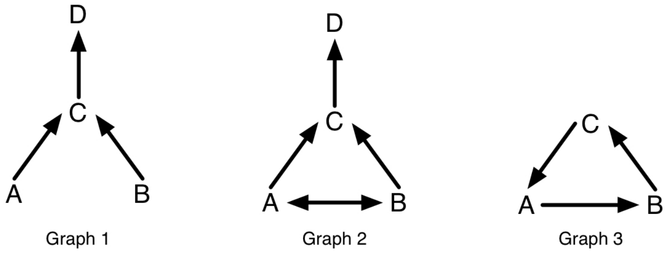


Figure 2.7: Three argument graphs in which letters represent arguments and arrows are defeat relationships between the arguments.

The foundations of Dung's approach are based on the concepts of: reinstatement;

acceptance of an argument with respect to a set of arguments; conflict-free sets; and, admissible sets.

Reinstatement captures the idea that a defeated argument can become undefeated if its attacker is defeated. In Graph 1 of Figure 2.7, argument D is defeated by C, but reinstated by A and B.

For Dung, acceptance is in relation to a set of arguments, so-called *collective acceptability* - others, such as Krause *et. al.* [74], calculate acceptance solely by considering the argument in question: *individual acceptability.*

An argument is *acceptable with respect to a set of arguments* if the argument is not attacked, or is reinstated, by a member of the set. A *conflict-free* set of arguments is one where no member of the set attacks another member. If each member of a set is acceptable with respect to the set, and the set is conflict-free, Dung calls it an *admissible* set. This notion is used to define a series of semantics for determining the acceptability of an individual argument; being a member of an admissible set can be considered the weakest form of acceptability.

Grounded semantics

Grounded semantics is the simplest semantics proposed by Dung. The idea is to construct the maximal admissible set. This may be done by seeding the set with all arguments that are not attacked, and then recursively adding all arguments that are acceptable to that set until no more can be added or conflict is caused. For example, in Graph 1 of Figure 2.7 the set would be seeded with A and B. These arguments defeat C so D can be added. Thus the *grounded extension* is $\{A, B, D\}$.

The advantage of grounded semantics is that it always produces an extension. In Graph 2 of Figure 2.7 the set cannot be seeded because all the arguments are attacked, thus the extension is the empty set. Such behaviour is a consequence of the decision to create a single consistent set - doing so rules out possibly valid alternatives.

Similar approaches are used by, amongst others, Pollock [45].

Preferred semantics

Unique status assignment approaches, for example grounded semantics, adopt the idea that there can be only one solution, and thus one valid result. However, to solve the problem of Graph 2 in Figure 2.7 it is necessary to take the viewpoint that there may be multiple sets of possibly valid arguments, so-called *multiple status assignment*.

Preferred semantics aims to construct one or more maximally admissible sets with respect to set inclusion. Inconsistency is resolved by producing multiple sets, each of which is consistent. In Graph 2, one set contains A and a second B. As A defeats C the first *preferred extension* is $\{A, D\}$ and similarly the second is $\{B, D\}$.

With these semantics, it is possible to define a credulous reasoner as one who accepts an argument if it is in any extension, and a sceptical reasoner as one who only accepts an argument if it is in every set. Therefore, D is sceptically accepted, B is credulously accepted, and C is defeated.

Preferred semantics can still provide an empty set as the sole extension. Graph 3 of Figure 2.7 features a so-called *odd defeat loop* (as opposed to the *even defeat loop* in Graph 2). The cycle is problematic because the conflict cannot be settled by placing each conflicting argument in a separate extension. To illustrate, if A is in the first extension C should be added because A reinstates C. However, C defeats A meaning that the extension is inconsistent. In this example it is not possible to create a consistent maximally inclusive set with respect to set inclusion.

Another problem with preferred semantics is their computational overhead. Grounded semantics can be computed with a linear algorithm; however, Dimopoulos and Torres [82] show that determining if a set is a preferred extension is CO-NP-complete.

Stable Semantics

Dung [31] also produced *stable semantics*; which shall not be discussed further as it is of no relevance to this work.

2.5 Argument visualisation

As discussed by Reed *et. al.* [33], the representation of arguments in diagram form has a long history with early applications including the recording of legal evidence, e.g. *Wigmore diagrams* [83], and logic, e.g. the *Toulmin scheme* [70].

There are many different styles and visualisations of argument. This chapter shall explore some of them through the use of an artificial classification. Category one discusses systems that visualise arguments often providing their user with a mechanism to analyse the argument's form. The second class contains systems that include a collaborative element, and thus help a range of people work together to reach a common goal. Argument visualisations associated with systems that perform argumentation, e.g. argumentation systems, are reserved for the final class. Instances of the third group may not be implemented by their creators.

2.5.1 Visualisation tools

Tools such as Araucaria [66] provide a mechanism to visualise arguments that can help with their study and pedagogical efforts [84]. These tools allow the user to draw arguments on the screen identifying the component parts and the links between them, e.g. an argument in Araucaria is shown in Figure 2.8. There is a wide range of such tools including: *Reason!Able* [85] (superseded by a commercial product called Rationale[3]), *Belverdere* [86], and *Convince Me* [87].

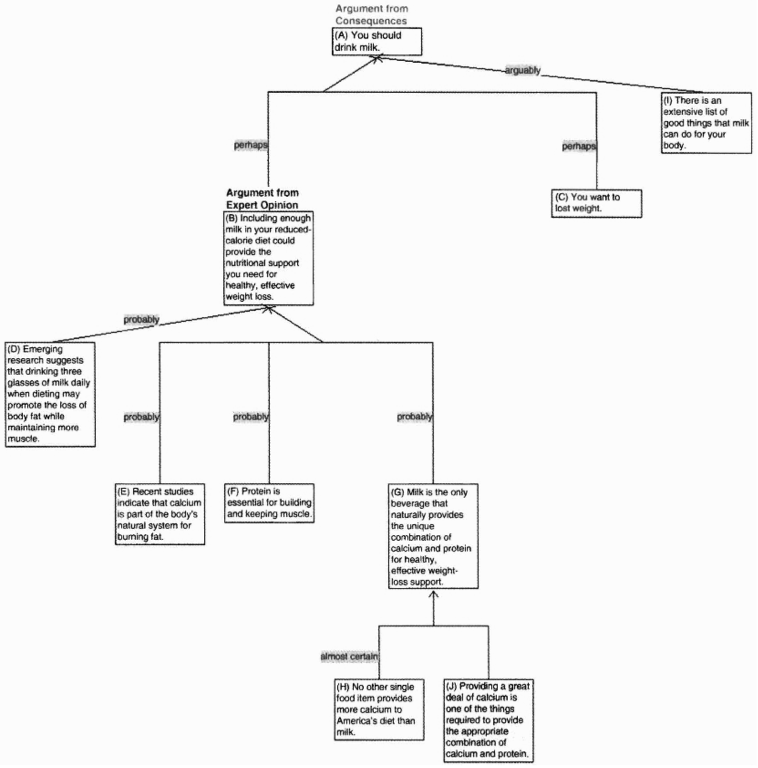
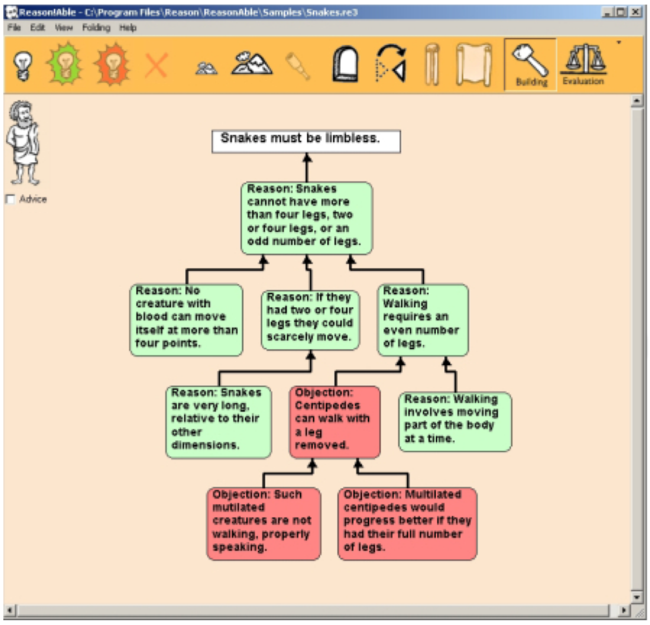

Argument from Consequences
(A) You should drink milk.
arguably
(I) There is an extensive list of good things that milk can do for your body.
perhaps
perhaps
Argument from Expert Opinion
(B) Including enough milk in your reduced-calorie diet could provide the nutritional support you need for healthy, effective weight loss.
(C) You want to lost weight.
probably
(D) Emerging research suggests that drinking three glasses of milk daily when dieting may promote the loss of body fat while maintaining more muscle.
probably
probably
probably
(E) Recent studies indicate that calcium is part of the body's natural system for burning fat.
(F) Protein is essential for building and keeping muscle.
(G) Milk is the only beverage that naturally provides the unique combination of calcium and protein for healthy, effective weight-loss support.
almost certain
(H) No other single food item provides more calcium to America's diet than milk.
(J) Providing a great deal of calcium is one of the things required to provide the appropriate combination of calcium and protein.

Figure 2.8: Araucaria argument visualisation - Figure 1 from [33].

[3] http://austhink.com


ReasonIAble - C:\Program Files\Reason\ReasonAble\Samples\Snakes.re3
File Edit View Folding Help
Building
Evaluation
□ Advice

Snakes must be limbless.

Reason: Snakes cannot have more than four legs, two or four legs, or an odd number of legs.

Reason: No creature with blood can move itself at more than four points.

Reason: If they had two or four legs they could scarcely move.

Reason: Walking requires an even number of legs.

Reason: Snakes are very long, relative to their other dimensions.

Objection: Centipedes can walk with a leg removed.

Reason: Walking involves moving part of the body at a time.

Objection: Such mutilated creatures are not walking, properly speaking.

Objection: Multilated centipedes would progress better if they had their full number of legs.

Figure 2.9: Reason!Able argument visualisation from [85].

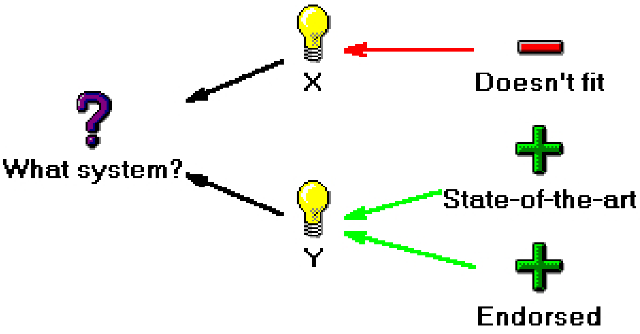


Figure 2.10: An IBIS argument map - Figure 7 from [88].

In general, as their default style, these tools provide a standard tree-like graph with the conclusion at the top, e.g. Figure 2.9. Premises are boxes linked to the conclusion (also a box) by directed arrows. Conclusions are pointed to, and the source of the arrow is a premise. Frequently, extra customisations are available. For example, Reason!Able allows the user to flip the diagram so the conclusion is at the bottom, left, or right. Araucaria takes this further by providing alternative visualisation models, in particular a Toulmin scheme [70] and Wigmore diagram [83]. Belvedere and Convince Me allow their users to drag and drop the boxes, enabling them to create any presentation style they wish. Some systems, e.g. Araucaria, provide the ability to customise the arrows. This can be seen in Figure 2.8 where a text label indicates which scheme the argument is an instance of (i.e. argument from consequences [2]).

2.5.2 Collaborative work tools

There are a number of tools designed to help humans work together to reach a common goal, for example make a business decision. Some of these tools are built upon underlying argumentation [89], e.g. those that implement IBIS [88]. IBIS, or *issue-based information system*, is a mechanism to help solve so-called *wicked problems* [90]: problems that cannot be tackled with standard scientific problem solving approaches. IBIS based approaches have a common graphical presentation of information, Figure 2.10 presents a simple example taken from Conklin [88]. In IBIS icons are used to add meaning to the basic graph form. A *question mark* icon indicates the issue to be discussed. The *light bulbs* represent possible solutions to the issue. A *green plus* is an argument supporting one solution, with a *red minus* being an argument against a solution. Thus the example in Figure 2.10 presents arguments for choosing system Y over system X.

One of the first systems to implement IBIS was gIBIS [91]. Both IBIS tools and the method have evolved. IBIS is now used as the heart of *dialogue mapping* - a method for having successful group communications that treats IBIS as a grammar on which it can build [92]. *Compendium* [93] represents the current state of the art.

In addition to standalone tools web-based mechanisms for collaboration are becoming common. Since inception the web has been used as a mechanism to help humans communicate and collaborate; now this communication can be structured to improve the quality and fairness of any debate. One site attempting to do this is *DebateGraph*[4]. It is similar to IBIS in that it commences with an issue that people wish to discuss: the issue is in the centre of the visualisation (called a *map*).

A series of maps taken from the DebateGraph site are presented in Figure 2.11. In part A the question asked is, *Should Iceland join the EU?*. This is related to the debates on *Iceland and the economic crisis*. There are two viewpoints. Examining

[4] http://debategraph.org

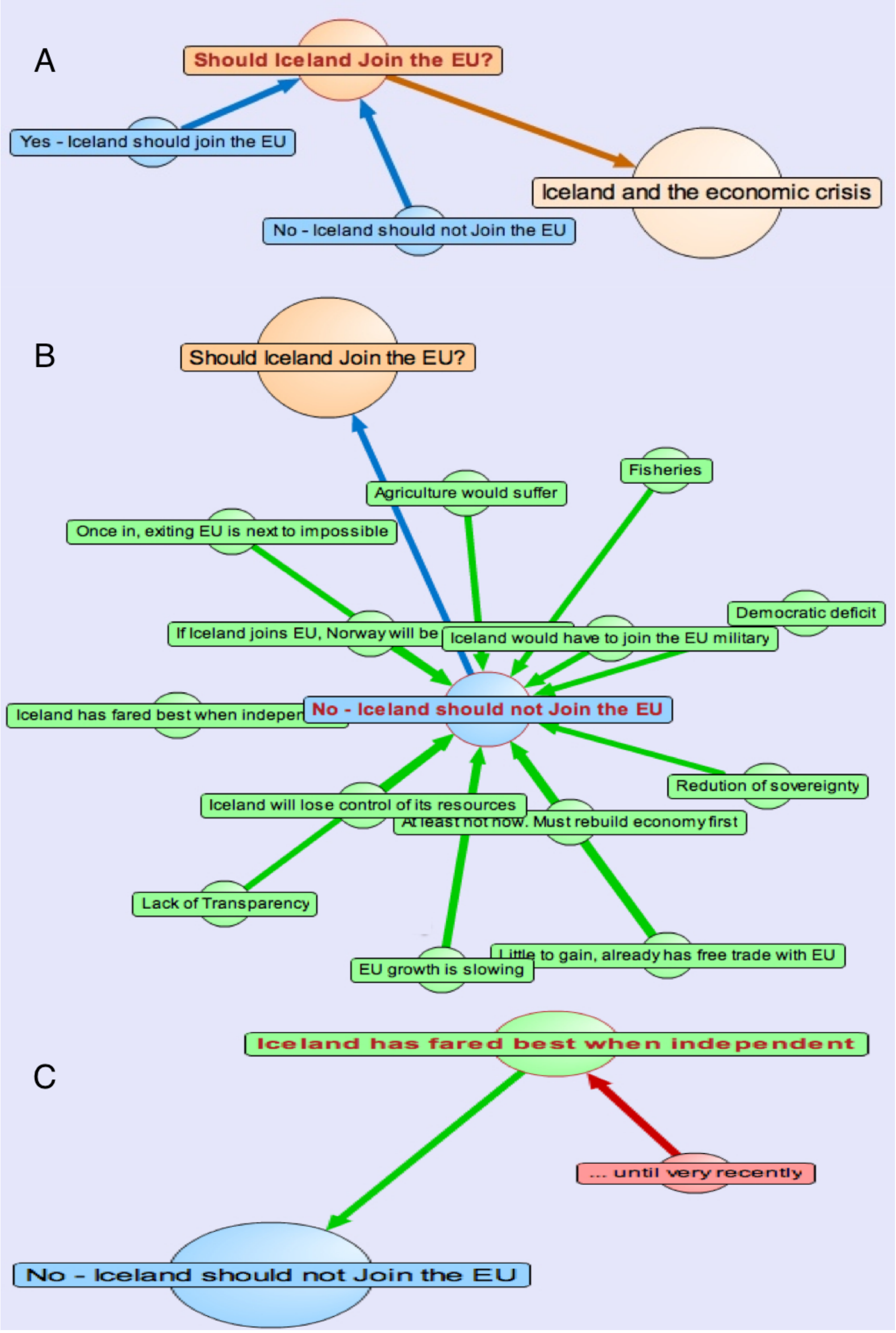

A
Should Iceland Join the EU?
Yes - Iceland should join the EU
Iceland and the economic crisis
No - Iceland should not Join the EU
B
Should Iceland Join the EU?
Fisheries
Agriculture would suffer
Once in, exiting EU is next to impossible
Democratic deficit
If Iceland joins EU, Norway will be
Iceland would have to join the EU military
Iceland has fared best when indepen
No - Iceland should not Join the EU
Redution of sovereignty
Iceland will lose control of its resources
At least not now. Must rebuild economy first
Lack of Transparency
Little to gain, already has free trade with EU
EU growth is slowing
C
Iceland has fared best when independent
... until very recently
No - Iceland should not Join the EU

Figure 2.11: Argument map from DebateGraph.

the *no* view, reveals Figure 2.11 part B. Now *No* becomes the central issue, which is related to the debate *Should Iceland join the EU?.* Various reasons why Iceland should not join are given in green nodes. Clicking on *Iceland has fared best when independent* brings Figure 2.11 part C, where the argument is attacked by the red node, which is presumably referring to the economic problems of 2008/9.

Colours are used in the graphs to identify the role of the nodes, e.g. green for support, and red for attack. By linking the different levels of a debate, and debates on related topics, the maps provide a mechanism to walk-though debates, focusing on individual aspects of interest. The underlying debates, from which the maps are generated, are conducted by the site's user community - they create topics and add their opinions, other users rate these and add their own views. The maps represent one way in which the site presents a summary of this information to its users, the other being a simple hierarchical tree.

2.5.3 Visualisations from argumentation systems

The third category of argument visualisations is related to systems that actually perform argumentation. Diagrams can be used by these systems to present the arguments they have created, or they can be used by the creators of such systems to help clarify their analysis of arguments.

The philosopher Pollock is an example of the latter. During his career he used a range of diagramming styles and techniques for the presentation of examples. This ranged from simple diagrams largely drawn using text-based characters (e.g. see Figure 2.12) to sophisticated diagrams involving colour and different types of line (e.g. Figure 2.13). Throughout his work the basic ideas and elements of the presentation are relatively constant: diagrams are kept simple; arguments can be either horizontal (with the conclusion at the right) or vertical (with the conclusion at the bottom); and, they often feature embedded logic. This use of logic is in stark contrast to the natural language used in the systems discussed in the previous two categories. The above systems are attempting to deal with the real world and thus focus on real world arguments using natural language, whereas Pollock is trying to formalise defeasible reasoning[5] accordingly his graphs are focused in a logical world.

One system for performing argumentation that presents the resulting arguments visually is *ArguMed* [94]. This employs a simple tree-like graph; yet, it is different in style to those proffered by Araucaria and Reason!Able. In this case, see Figure 2.14, the propositions are all directly under the conclusion. Different styles of arrowhead are used to clarify the relationship of the proposition to the rest of the argument, for example a cross is used to identify an undercut (see Section 2.4.4 for more information on undercuts).

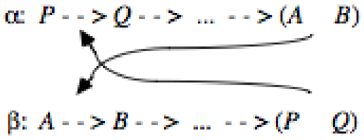

α: P - - > Q - - > ... - - > (A B)
β: A - - > B - - > ... - - > (P Q)

Figure 2.12: A simple argument graph drawn by Pollock - Figure 2 in [45].

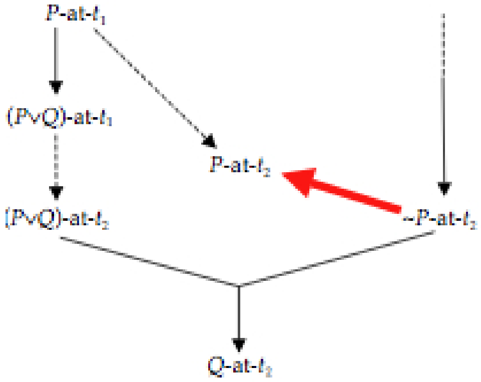


Figure 2.13: A sophisticated argument graph drawn by Pollock - Figure 9 from [53].

Other systems for argumentation stick to the normal tree-like graph, for example the Carneades Argumentation Framework [95] (a typical argument can be seen in Figure 2.15) - in which a range of arrows and arrowheads are employed to convey the meaning of the relationship between the boxes. Carneades can help users create and evaluate arguments diagrammatically. A similar task is performed by *AVER* [96], although it concentrates on helping crime scene investigators explore gathered evidence and inferences made; it too uses the common tree-based paradigm for presentation.

[5] The term used by philosophers to describe non-monontonic reasoning.

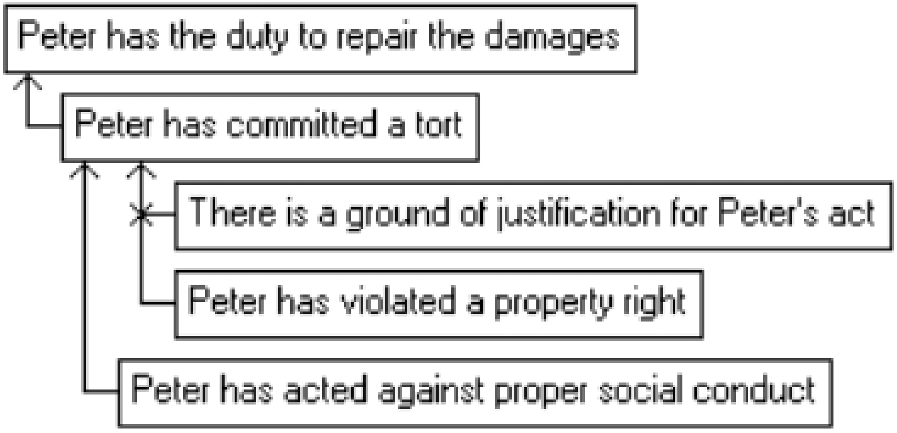

Peter has the duty to repair the damages
Peter has committed a tort
There is a ground of justification for Peter's act
Peter has violated a property right
Peter has acted against proper social conduct

Figure 2.14: ArguMed argument diagram - Figure 2 from [94].

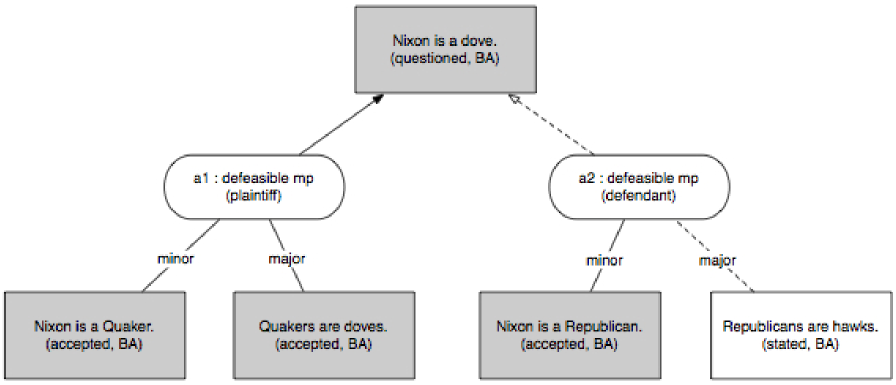


Figure 2.15: Carneades argument graph - Figure 1 from [95].

2.6 Applications of argumentation

Argumentation is applied in a range of areas, during this section some of the applications of argumentation in AI will be reviewed.

Reed and Grasso [25] suggest that the role of argumentation in AI may be divided roughly into two categories: modelling with argument, and modelling of argument. The latter class employs argumentation theory to model natural language arguments that take place between humans, for example a legal debate (see Section 2.6.1 for more information). The former captures the idea of using argumentation to tackle problems like non-monotonic reasoning, i.e. areas that do not naturally contain arguments.

As NMR has been discussed in Section 2.2, this section will start with an exploration of AI & law, before progressing to more biologically relevant domains such as medicine (in Section 2.6.2) and science (Section 2.6.3).

2.6.1 AI & law

AI & law is one of the most active domains for argumentation research. Because law involves a series of arguments and persuasion dialogues, it provides an excellent playground for researchers [5]. Nevertheless, not all work in AI & law uses argumentation theory [97]. However, work has focused increasingly on the natural link between the practise of law and argumentation theory.

The idea of studying jurisprudence to improve the understanding of arguments first originated from Stephen Toulmin [70]. He presented a model of argument (the Toulmin scheme) that provides a useful model for interpreting and understanding natural language arguments, be they legal or otherwise.

One of the first argumentation-based systems was *HYPO* [98]. It models *case-based* reasoning (the application of *precedence*) which is central to U.S. law. HYPO is notable for the application of arguments within a dialogue. HYPO evolved into *CATO* [99] a system used to teach precedence based reasoning to legal students.

HYPO and CATO apply an informal theory of the method used to conduct legal reasoning; yet, formal accounts of reasoning exist. To handle the defeasible aspects of legal reasoning non-monotonic approaches were investigated, with argumentation theory providing one of the most popular. Dung's framework [31] was first applied by Prakken and Sartor [71] to investigate the possibility of arguing over the preference order assigned to arguments. Dung's work has been applied to case-based reasoning too, e.g. Prakken [100].

Gordon [101] produced a system, *The Pleadings Game*, which wraps a civil law dispute in a dialectical process. Humans create arguments, the computer evaluates them, and controls the game ensuring it is fair. In this way, access to an argumentation system is controlled through the dialogue. Evolutions of this work include *TDG* [102] and *dialaw* [103]. TDG uses the Toulmin scheme to improve the readability of its arguments.

With the application of a game, tactics become germane. The player who is to make the next move may be able to use any one of multiple arguments, so which argument should he/she/it[6] choose? This work was started by *CABARET* [104], a system that attempted to form strategies by classifying arguments and determining how/when they could be best used or attacked. The classification of arguments according to their form is continued under the banner of argumentation schemes.

The most famous scheme is that created by Toulmin and discussed above. It is used in dialogue games, e.g. TDG [102], and tools that visualise arguments, e.g. PLAID [105]. Toulmin's scheme is generic. Others, e.g. Walton [55], try to create specific schemes for specific types of argument some of which, e.g. the *argument from expert opinion* [106], directly target applications in law. Furthermore, Verheij approaches his work, e.g. [94, 48, 62], through law.

AI & law was one of the first domains to investigate the visualisation of arguments with Wigmore diagrams in the 1930s [83]. They deliver a mechanism to analyse all the information in a case, record the relationships (support, attack) between the information, and assign each piece of information a degree of belief (strength). Wigmore diagrams are used in the *MARSHALPLAN* system [107], which is designed to help visualise preliminary fact investigation. Subsequent work evolves the diagramming techniques, and now may implement some of Walton's schemes, e.g. AVER [96]. AVER, a tool for crime scene investigators, goes beyond diagramming and provides methods to evaluate the argument displayed. More generic tools can be applied for diagramming, e.g. Araucaria [66], or diagramming and evaluating legal arguments, e.g. Carneades [95].

In summary, the domain of AI & law has been central to the development of computational argumentation and very influential in the wider world of argumentation theory.

2.6.2 Medicine

Medicine employs argumentation theory primarily in two main ways. The first is the application of argumentation for decision support and explanation. The second is the generation of natural dialogue for use in communication with the user.

[6] The player may be a software agent.

One of the earlier instances of the latter category is from Grasso *et. al.* [11]; the authors avail themselves of schemes to generate arguments as part of a persuasive discussion in which the system tries to improve a human user's eating habits. An obvious descendent of this work is by Mazzotta and de Rossi [108]. It exploits Walton's notion of schemes to generate persuasive emotional arguments to convince someone to improve their diet. Related work by Alberg [109] tries to deal with malnutrition in elderly people. Day [110] applies similar ideas to try and persuade someone to improve their health by exercising.

Bickmore and Sidner [111] combine the ideas of communication and decision support to propose a system that takes advantage of a dialogue framework to improve communication between medics and patients resulting in an improved treatment plan. Schultz and Rubinelli [112] aim for a similar outcome; however, they advocate starting by trying to understand the communication mechanism between medics and patients. To this end they apply IL techniques to a set of documented medic-patient consultations.

Hunt *et. al.* [113] show that decision-support systems in medicine can be very helpful for medics. The exceptions are systems targeted at diagnosis which universally fail. Examples of decision support include the work of the *advanced computation laboratory* at *cancer research UK* (now COSSAC[7]). They commenced with the creation of an argument-based logic for reasoning under uncertainty, the *logic of argumentation* (LoA) [74]. This is employed within a model of clinical guidelines, PRO*forma* [114], which allows the guidelines to capture reasons for and against performing a particular treatment. PRO*forma* is utilised within a series of tools such as *REACT* [115] (decision support for medical planning by a single medic), and *LISA* [116] (a clinical information and decision support system for Leukaemia treatment). These systems, and others, have been evaluated and shown to be successful [117].

In addition to work using LoA, COSSAC built an implementation of a Dung-style framework through their involvement in the *ASPIC* project [7]. Williams and Williamson [81] use the Dung inspired framework of Prakken and Sartor [71] to generate arguments that act as explanations of reasoning performed by other means.

Other research teams have brought argumentation to bear in order to provide

[7] `www.cossac.org`

explanations for clinical decision support systems. Shankar *et. al.* [118] operate the Toulmin model of argument to generate arguments that explain a decision made by the *athena* decision support system [119]. In Tolchinsky *et. al.* [120] a framework utilising argumentation to decide on the viability of a human organ for transplant is discussed. Many agents (software and human) create arguments, which are sent to, and evaluated by, a mediator agent. The mediator uses argumentation to make a final decision: critical questions are employed to assess the arguments sent, then an implementation of Dung's framework evaluates the status of those arguments.

Part of the reasoning undertaken by Tolchinsky *et. al.* [120] is case-based, which provides an overlap with the legal domain. Dolins and Kero [121] suggest that this form of reasoning is important for medicine, and next-generation medical informatics systems. Another common link between law and medicine is the Toulmin scheme, which is exploited by both communities to represent arguments in a natural way. An example from medicine is Green [12], where it forms part of a natural language generation system designed to help create pamphlets for patients engaging in genetic counselling.

2.6.3 Science

Despite the prominence of argumentation in medical informatics, it is still relatively untried in main stream science (chemistry, biology, and physics). One noticeable exception from biology is the work of Jeffreys *et. al.* [9]. It employs a simple model of argumentation to evaluate the output of a bioinformatics tool. Their work shows that argumentation is as effective as other mechanisms such as decision trees. Furthermore, Jeffreys *et. al.* intimate that argumentation frameworks and Bayesian networks are not directly comparable because they are designed to tackle different problems.

Argumentation's relationship to science seems to be mainly in the field of pedagogy, where it is used to help students learn to create good scientific arguments, e.g. McNeil and Pimentel [122]. As such, the natural language aspects of the field are used, as opposed to the computational elements focused on in this thesis.

It should be remembered that argumentation theory has roots in the fields of psychology, e.g. Voss and Van Dyke [123], and cognitive science, e.g. Erduran and Jiménez-Alexiandre [124]. Consequently these disciplines both use and contribute to the study of argumentation theory.

2.6.4 General applications

Throughout this chapter a number of application areas in addition to medicine, law, and science have been discussed. These include: agent communication, e.g. Rahwan *et. al.* [8]; collaborative working, e.g. de Moor and Aakhus [34]; and pedagogy, e.g. McNeil and Pimentel [122]. Furthermore, argumentation is being applied to related areas such as e-democracy, e.g. Cartwright and Atkinson [125]; practical reasoning, e.g. Gilbert [126]; and decision support, e.g. Bury *et. al.* [116]. In reality, argumentation may be applied to any area in which information is uncertain, because in such situations there is room for debate. As an example of how far argumentation has spread, Trojahn *et. al.* [6] demonstrate that it can be used to combine different techniques for ontology matching.

2.7 Summary

This chapter reviews the world of argumentation research. The primary aim of this chapter is to furnish the user with the knowledge necessary for an understanding of the rest of this document.

Whilst this chapter undoubtably focuses on an AI-centric view of argumentation, brief consideration is given to some philosophical aspects - for example the purpose of argumentation schemes, and the ways in which these schemes can be verified. The role of other fields in developing argumentation theory is acknowledged, but not dwelt upon. Similarly, the application of argumentation related technologies in these fields is not considered.

Subsequent chapters use a third party argumentation toolkit to perform computational argumentation. The toolkit is based on the notion of an *argumentation framework*, in which an implementation of Dung's seminal work [31] analyses whether arguments are true or false. Hence particular attention is paid to an exploration of frameworks and an explanation of Dung's contribution.

Walton's [55] notion of *argumentation schemes* - loosely defined as a natural language template for documenting arguments - plays a pivotal role in future chapters.

Likewise, Verheij's [62] approach for converting schemes into formal logic inference rules. Both topics are discussed in depth.

A variety of research and commercial activity has considered the best approaches for visualising arguments. These efforts are reviewed before applications of computational argumentation are considered within the domains of AI & law, medical informatics and general science.

Chapter 3

Biology

The biology behind this document's use case is discussed, prior to outlining the incentive for the research in Chapter 4. An introduction to some basic biological concepts is found in Section 3.1. Subsequently the use case's domain is explored beginning with the notion of gene expression (Section 3.2), and continuing with the developmental mouse (Section 3.3). Following on to that a brief review of the main forms of gene expression experiment - including *in situ* hybridisation - is conducted in Section 3.4. Before the chapter concludes in Section 3.6, Section 3.5 examines some of the key resources in this field focusing upon those utilised in this thesis.

3.1 Biology 101

To commence, a brief overview of basic biology is provided.

Species is the name given to a group of similar individuals that are members of the same close biological family. Constituents of the same species should be able to reproduce as should their offspring. An individual member of a species is called an *organism*. Organisms are comprised of *systems*, e.g. the respiratory system. These systems contain *organs* (for example the lungs), which are in turn composed of similar groups of *cells* known as *tissues*. Cells are the basic unit of life and contain distinct structures (so-called *organelles*), e.g. the nucleus.

In mammalian cells the nucleus contains *chromosomes* - a chromosome is a long DNA (*deoxyribonucleic acid*) molecule. The chromosome may be split into a series of small units called *genes*. Each gene supplies the information required by the cell to function or develop. Every cell in an organism contains the same DNA; however, for an individual gene to be used by a cell it must be “switched on”. Genes that are switched on are said to be *expressed.* Similarly, genes that are not switched on are *not expressed.*

Genes enact changes in a cell by causing the formation of *proteins.* In addition to carrying material out of a cell and around the body, proteins are molecules that alter the behaviour and function of the cell. Consequently, proteins have a wide range of activities ranging from catalysing chemical reactions, and switching off genes, to being the building blocks used to support cells and create structures such as hair.

Gene regulation is the process of controlling the expression level of a gene. Regulation is often a secondary result of a complex chemical reaction. During the reaction proteins are generated that either reduce or increase the current level of expression. If reduced sufficiently, the gene will be “switched off”.

Unwanted or abnormal features such as cleft lips are a result of certain genes having the “wrong” level of expression, e.g. genes that should be expressed are not expressed. To gain an understanding of these features, the set of genes expressed in normal healthy tissues must be compared to the set of genes expressed in abnormal or unhealthy tissues. The difference between the two sets provides an indication of the root cause of the abnormality and a basis for further research into a cure or prevention.

A more in-depth discussion on the basics of biology is provided in standard textbooks such as Campbell *et. al.* [127].

3.2 DNA, genes and gene expression

As mentioned in Section 3.1, genes are instructions that control the development of an organism by affecting the type and number of proteins produced at any one time.

Genes are small units, comprised of DNA, found in the chromosome of every cell. Initially, the DNA is *transcribed* to form RNA (*ribonucleic acid*). The RNA is then *translated* into proteins. When a protein has been produced, the gene can be said to be expressed.

A series of *nucleotides* constitute a DNA molecule. There are four distinct nucleotides found in DNA: *adenine*, *cytosine*, *guanine*, and *thymine.* Each nucleotide is a short molecule that acts as a "link in the chain" of the longer DNA molecule. The distribution of these nucleotides varies throughout the length of the molecule creating the so-called *genetic sequence*. It is this variation that encodes the information stored in the DNA.

DNA is broken into a series of triplets called *codons*; each is a sequence of three nucleotides. Particular codons demarcate the individual genes within the long DNA molecule (i.e. the chromosome). Different codons are used to split genes into segments. A *promoter* is an area that controls the regulation of the gene; depending on which protein binds to the promoter, the gene will be ignored or transcribed. The sequence of nucleotides to be transcribed into RNA is likewise a region wrapped by codons.

Gene expression experiments are interested in determining which genes are active (expressed) in a specific location within a particular organism at a precise time. Measuring the type and number of proteins present gives a true picture of which genes are expressed. Not every RNA is translated into a protein, so looking at RNA levels merely allows an estimate of the gene expression to be made. Examining DNA indicates whether or not a gene is present - it conveys nothing about the expression level of that gene.

In this document's use case, gene expression is studied with respect to the developmental mouse.

3.3 The developmental mouse

It is not possible to experiment on humans for practical, ethical, and legal reasons. Consequently substitute organisms are used by scientists. These are known as *model organisms*. A wide range of plants, animals, fish, and insects are studied. This work concentrates on one of those, the mouse.

"Mouse" is the common name for the animal with the latin name *Mus*. There are many species ranging from *mus musculus* (the house mouse) to *peromyscus maniculatus* (the deer mouse). In addition to knowing the species of mouse used, it is necessary to know whether or not the mouse is a *mutant*. So-called *wild type* mice have a normal, or natural, set of chromosomes. Whereas mutants are bred to insure they have a particular trait, for example cancer. To enable the comparison of results obtained from different individuals, the mice must be genetically uniform, i.e. the mice must be of the same *strain*.

The mouse *develops* from a single cell into a mammal with a complex anatomy comprising countless cells. This development was studied by Karl Theiler [128]. He split the development of the house mouse into 28 distinct stages, called *Theiler Stages*. The first 26 Theiler Stages deal with the unborn mouse. The final two describe the newborn and then postnatal adult mouse. Current convention creates a split between the unborn mouse and the final two stages. The former group being called the *developmental mouse* and the latter the *adult mouse*.

A Theiler Stage is accompanied by an approximate time since conception measured in days, called *days post conception* (dpc). Additionally, it includes a description of the anatomy at that stage, and highlights what has changed from the previous stage. Diagrams illustrate these differences. A summary and outline of these stages can be found at the Edinburgh Mouse Atlas Project (EMAP) website[1].

Each Theiler Stage has an associated anatomy, and corresponding ontological representation - see Figure 3.1. Although multiple anatomies exist, the main anatomy in this work is EMAP.

3.3.1 EMAP anatomy

The acronym EMAP is ambiguous because it is overloaded. It applies to both the *Edinburgh Mouse Atlas Project*[2] and the anatomy developed as one part of that project. Part of the anatomy ontology for TS 14 can be seen in Figure 3.2. The anatomy of the developmental mouse is described using a series of *part of* relations, thus looking at Figure 3.2 it is obvious that the *future brain* is part of the *central nervous system*, which is in turn part of the *nervous system* and that is part of the *organ system*.

Each structure in the ontology is given a unique identifier in the form *EMAP:number*, e.g. EMAP:152. Moreover the structure has a name, for example *future brain*. This is the structure's *short name*. Its *full name* is the entire path from the root node of the ontology to its short name, e.g. mouse > embryo > ectoderm >

[1] `www.emouseatlas.org/emap/ema/theiler_stages/StageDefinition/stagedefinition.html`

[2] `www.emouseatlas.org`

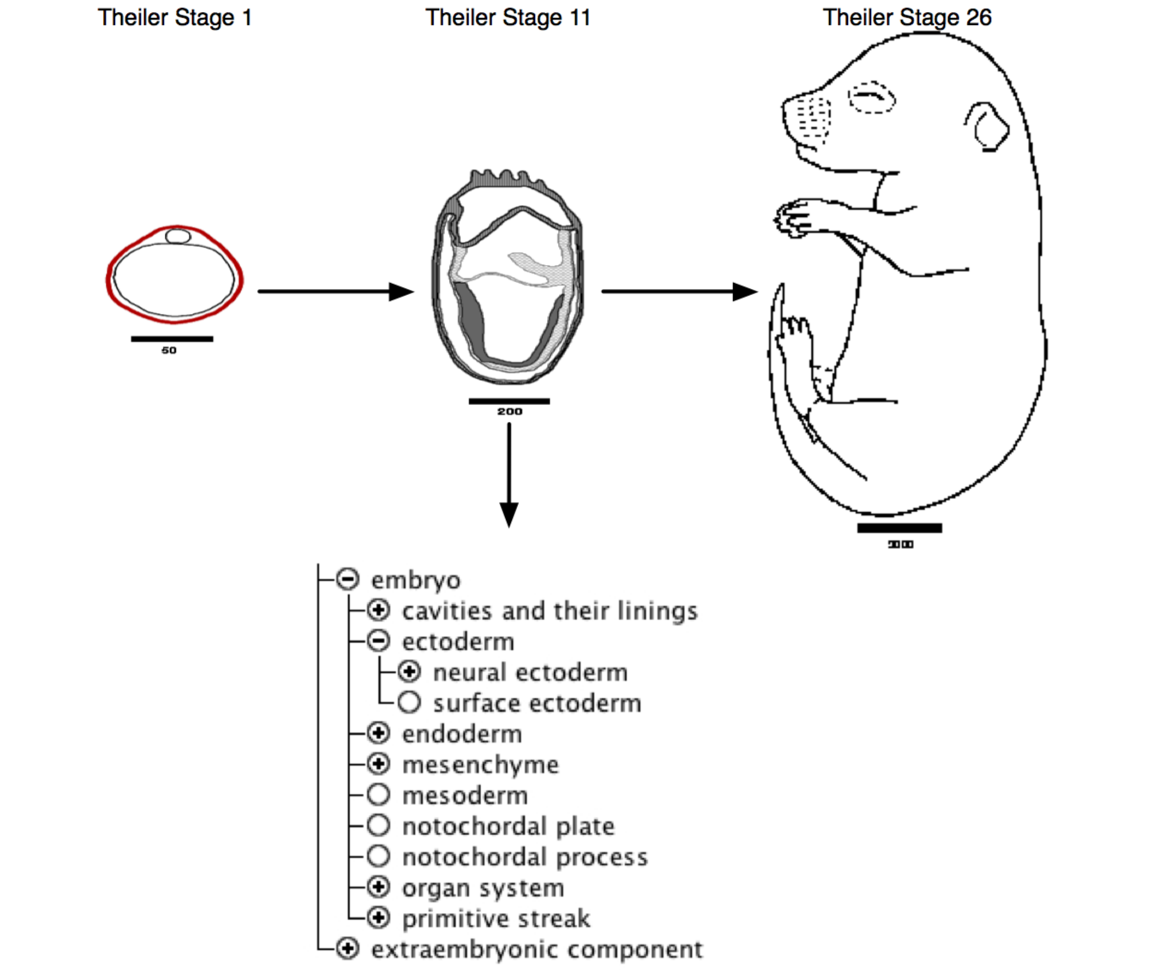
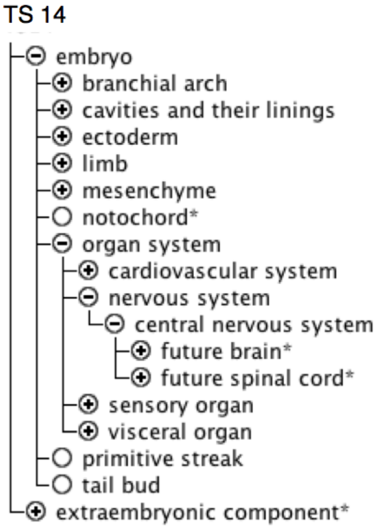


Figure 3.1: Illustrating the development captured by the different Theiler Stages; each stage has an anatomical and associated ontological representation. The anatomy ontology shown here is a subset of EMAP Theiler Stage 11.

neural ectoderm > future brain.

The same structure can appear in multiple stages, and can have the same short and full names in these stages; the above example applies equally to Theiler Stages 11, 12, and 13. The unique feature of the future brain in these different stages is its ID: in TS11 it has EMAP:152; in TS12 EMAP:235; and in TS13 EMAP:441.

More details on the EMAP anatomy (ontology) can be found in Baldock and Davidson [129].



```
TS 14
├⊖ embryo
│ ├⊕ branchial arch
│ ├⊕ cavities and their linings
│ ├⊕ ectoderm
│ ├⊕ limb
│ ├⊕ mesenchyme
│ ├○ notochord*
│ ├⊖ organ system
│ │ ├⊕ cardiovascular system
│ │ ├⊖ nervous system
│ │ │ └⊖ central nervous system
│ │ │   ├⊕ future brain*
│ │ │   └⊕ future spinal cord*
│ │ ├⊕ sensory organ
│ │ └⊕ visceral organ
│ ├○ primitive streak
│ └○ tail bud
└⊕ extraembryonic component*
```

Figure 3.2: Part of the EMAP anatomy ontology for Theiler Stage 14.

3.4 Gene expression experiments

There is a wide range of techniques to determine the expression level of a gene. These mechanisms differ not only in their method, but in their precise focus. Primarily, experiments can concentrate on the location of expression or on the quantity of the gene expressed.

Often experiments rely on the close association between individual genes, RNA, and proteins - a gene is transcribed into RNA, which may be translated into a protein. This means a protein or RNA can be mapped onto a corresponding gene. Therefore, gene expression can be evaluated by examining the proteins or RNA contained in a sample.

One key principle relied on in many of the different experimental techniques is the notion of *hybridisation*. This is the idea of chemically bonding a *probe* to DNA, RNA or a protein. Normally probes are designed to bond with a single gene/RNA/protein, but this is not always the case. Probes are very conspicuous - for example they may be highly coloured or radioactive - the visibility of the probe provides an insight into where genes are expressed and the quantity of the expression found in that area.

A brief overview of several available techniques follows in Section 3.4.2. However, the technique of *in situ* hybridisation is given special consideration in Section 3.4.1, as it is the main class of experiment in this thesis's use case.

3.4.1 *In situ* hybridisation gene expression experiments

In situ is the latin for "in place". Accordingly *in situ* experiments focus on identifying precisely *where* a gene is expressed. They produce images that are either an entire mouse embryo (a so-called *wholemount*), or a slice/section (e.g. Figure 3.3) of that mouse. The areas of intense colour indicate where the gene is expressed. To document the link between the genes and the location of their expression the visual result may be mapped manually to an anatomy. The anatomy may have an ontological representation, e.g. EMAP, or a 3D spatial depiction (see Figure 3.4).

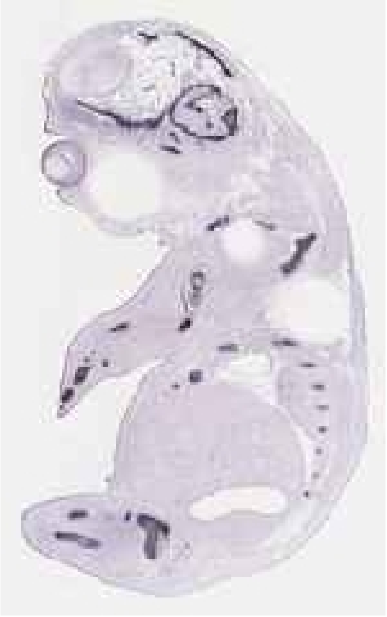


Figure 3.3: Result of an *in situ* gene expression experiment - EMAGE:6089.

3.4.2 Common experimental techniques

Gene expression techniques attempt to measure the level of gene expression by examining the RNA, or proteins present in a biological sample. In addition, DNA can be examined to determine if a gene is present. There are a variety of techniques designed to do these tasks including:

Southern blot A simple method of deciding if a known gene is present in a sample by looking at the DNA. A small sample is fixed onto a surface, and a probe washed over it.

Northern blot Similar to the southern blot, but this time measuring the RNA levels, and thus the level of gene expression.

Western blot Similar to the Northern blot, but this time measuring protein levels.

***In situ* hybridisation against mRNA** An *in situ* experiment, as described above, focusing on one type of RNA. It differs from the blots, because the sample is the surface and the probe is washed directly over it.

Immunohistochemistry A method of identifying a particular cell by determining which proteins are located inside it.

***In situ* reporter** A method of identifying gene expression by replacing a particular gene with a mutated form that may be detected easily, allowing the location of the gene to be tracked as the organism develops.

RT-PCR *polymerase chain reaction* (PCR) involves increasing the level of a particular RNA to make it more visible and thus easier to detect and count. Once the count has been performed, the amount of the original RNA can be calculated. RT stands for *reverse transcription*, thus in this particular form of PCR, the RNA is reverse transcribed to generate DNA, and then amplified.

RNase protection *ribonuclease* (RNase) *protection* provides a mechanism for detecting and quantifying specific RNAs within a sample. A series of probes are hybridised with the sample. The hybrid, formed by the reaction between a probe and the RNA, is separated from the remaining unused probe and sample. RNase is applied to the hybrid, it digests all the RNA leaving behind the successful probes. Determining which probes are left, and how many of each probe type is present, furnishes information on which genes are expressed and the quantity in which they are expressed.

SAGE *serial analysis of gene expression* is a technique to quantify the number of each gene expressed in a subject using RNA. The basic process involves extracting RNA from the cells in the subject. The RNA is converted into short sequences of DNA called *tags*. Each tag uniquely represents the gene from which the original RNA was translated. The tags are collected together and processed by a computer, which counts the number of each tag. The quantity of each tag indicates the quantity of the original gene that is expressed in the subject.

Microarray A technique to quantify gene expression. A microarray is a small chemically stable platform, which contains thousands of probes - each detecting a different gene. A sample is produced by chopping up and liquidising a piece of tissue from one or more organisms. The sample is washed over the platform, allowing the probes to hybridise. If a probe has hybridised with a RNA, that probe becomes visible. The greater the number of suitable RNA, the more hybridisation takes place. There is a direct relationship between hybridisation and visibility with increases in hybridisation leading to greater (more intense) visibility. The microarray is analysed by a computer, which determines the level of visibility of each probe, and quantifies the expression levels. Microarray experiments are very high throughput, which makes them popular. However, the fact that the biological sample has to be blended results in only coarse grain location information being obtained.

More details on gene expression and experimental techniques can be found in Avison [130].

3.5 Resources

The following section examines some of the main resources in the chosen field in addition to some of the large gene expression resources.

3.5.1 EMAGE

EMAP is the Edinburgh Mouse atlas Project. It is the umbrella name for a range of activities.

The first of these activities is EMA is the *Edinburgh Mouse Atlas.* EMA is responsible for the creation and maintenance of an anatomy (and corresponding ontology) for the developmental mouse. The ontology has the name EMAP, and was discussed in Section 3.3.1. In addition to the ontology, the project has produced a series of 3D computer models of the mouse, e.g. Figure 3.4. There is at least one 3D model for each Theiler Stage between 7 and 26. Each model comprises a number of *voxels* (volumetric pixels) that are stacked in a 3D space.

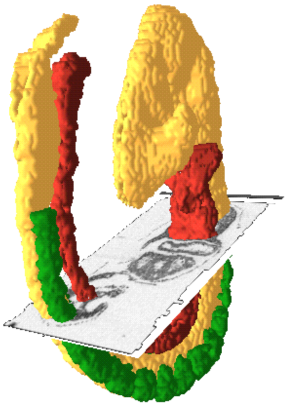


Figure 3.4: An illustration of the EMAGE 3D model for Theiler Stage 14.

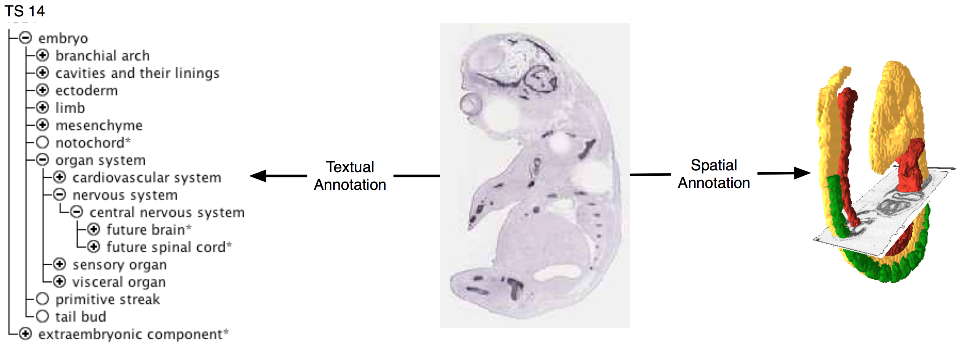


Figure 3.5: Textual annotations versus spatial annotations - spatial are linked to one of the 3D models whereas textual are tied to the anatomy ontology.

EMAGE[3] is the *Edinburgh Mouse Atlas of Gene Expression*, the second main activity of EMAP. This is a gene expression database that (re)publishes *in situ* gene expression data for the developmental mouse. When researchers perform an experiment they may publish it in a traditional journal, and then have it republished in EMAGE (or a similar resource). Alternatively, the researchers may submit their data directly to a resource, such as EMAGE, by-passing the orthodox scientific journals.

EMAGE publishes two forms of results (see Figure 3.5): those tied to the anatomy ontology (so-called *textual annotations*); and results linked to the 3D models (*spatial annotations*). EMAGE is unusual in this respect, because most gene expression re-

[3] www.emouseatlas.org/emage/

sources still do not use 3D models, and thus cannot produce spatial annotations.

EMAGE experimented briefly with a third type of annotation: textual annotations derived from spatial annotations. A computer algorithm was used to turn the spatial annotations into textual annotations. This proved to be expensive and unreliable. Consequently, no new annotations are being created; however, the set of generated annotations is still available.

In terms of content, EMAGE contains the following types of experiment: *in situ* hybridisation against mRNA, immunohistochemistry, and *in situ* reporter. In December 2011, it contained details of over 40,000 procedures.

More details on EMAGE can be found in Venkataraman *et. al.* [131].

EMAGE walkthrough

In order to illustrate the information supplied by a typical gene expression resource, a single experiment from EMAGE shall be examined. The experiment's web page can be seen in Figure 3.6; there is nothing remarkable about this particular page or experiment, it was selected because it is a typical example.

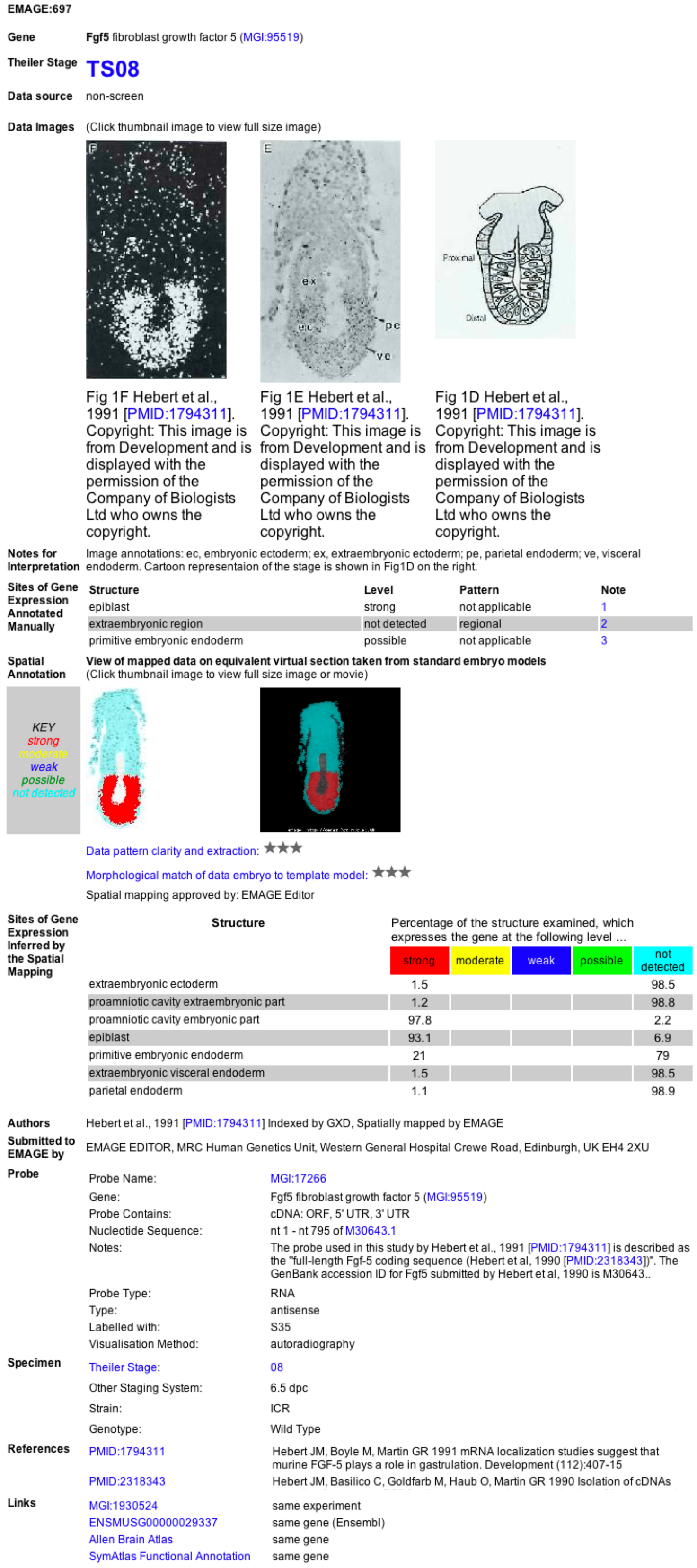
At the top left of the page, in bold font, is **EMAGE:697**. This is the accession identifier, a unique ID for the experiment.

Below that is information on the **Gene**. The gene's unique symbol (in this case *Fgf5*) precedes the gene's name (fibroblast growth factor 5). At the end of the line is a link to a page containing more details on this gene. That page is provided by a different resource, The *Mouse Genome Informatics* (MGI)[4]. The link is presented as the accession ID for the gene in that resource.

The next line indicates the **Theiler Stage** on which the experiment was performed, i.e. the age of the mouse when the experiment was conducted. Theiler Stage 8 corresponds to approximately 6.5 days after conception.

Data source indicates whether or not the experiment was performed by a screening program, in this case it was not. *Screening programs* are projects that are optimised for throughput and thus perform large volumes of experiments. The alternative is for the experiment to be conducted by a small research lab, possibly one specialising in the tissue or gene being experimented on.

[4] `www.informatics.jax.org/mgihome`



EMAGE:697

Gene	**Fgf5** fibroblast growth factor 5 (MGI:95519)
Theiler Stage	**TS08**
Data source	non-screen
Data Images	(Click thumbnail image to view full size image)

Fig 1F Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Fig 1E Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Fig 1D Hebert et al., 1991 [PMID:1794311]. Copyright: This image is from Development and is displayed with the permission of the Company of Biologists Ltd who owns the copyright.

Notes for Interpretation Image annotations: ec, embryonic ectoderm; ex, extraembryonic ectoderm; pe, parietal endoderm; ve, visceral endoderm. Cartoon representaion of the stage is shown in Fig1D on the right.

Sites of Gene Expression Annotated Manually

Structure	Level	Pattern	Note
epiblast	strong	not applicable	1
extraembryonic region	not detected	regional	2
primitive embryonic endoderm	possible	not applicable	3

Spatial Annotation

View of mapped data on equivalent virtual section taken from standard embryo models
(Click thumbnail image to view full size image or movie)

Data pattern clarity and extraction: ★★★

Morphological match of data embryo to template model: ★★★

Spatial mapping approved by: EMAGE Editor

Sites of Gene Expression Inferred by the Spatial Mapping

Structure	Percentage of the structure examined, which expresses the gene at the following level ...				
	strong	moderate	weak	possible	not detected
extraembryonic ectoderm	1.5				98.5
proamniotic cavity extraembryonic part	1.2				98.8
proamniotic cavity embryonic part	97.8				2.2
epiblast	93.1				6.9
primitive embryonic endoderm	21				79
extraembryonic visceral endoderm	1.5				98.5
parietal endoderm	1.1				98.9

Authors	Hebert et al., 1991 [PMID:1794311] Indexed by GXD, Spatially mapped by EMAGE
Submitted to EMAGE by	EMAGE EDITOR, MRC Human Genetics Unit, Western General Hospital Crewe Road, Edinburgh, UK EH4 2XU

Probe

Probe Name:	MGI:17266
Gene:	Fgf5 fibroblast growth factor 5 (MGI:95519)
Probe Contains:	cDNA: ORF, 5' UTR, 3' UTR
Nucleotide Sequence:	nt 1 - nt 795 of M30643.1
Notes:	The probe used in this study by Hebert et al., 1991 [PMID:1794311] is described as the "full-length Fgf-5 coding sequence (Hebert et al, 1990 [PMID:2318343])". The GenBank accession ID for Fgf5 submitted by Hebert et al, 1990 is M30643..
Probe Type:	RNA
Type:	antisense
Labelled with:	S35
Visualisation Method:	autoradiography

Specimen

Theiler Stage:	08
Other Staging System:	6.5 dpc
Strain:	ICR
Genotype:	Wild Type

References

PMID:1794311	Hebert JM, Boyle M, Martin GR 1991 mRNA localization studies suggest that murine FGF-5 plays a role in gastrulation. Development (112):407-15
PMID:2318343	Hebert JM, Basilico C, Goldfarb M, Haub O, Martin GR 1990 Isolation of cDNAs

Links

MGI:1930524	same experiment
ENSMUSG00000029337	same gene (Ensembl)
Allen Brain Atlas	same gene
SymAtlas Functional Annotation	same gene

Figure 3.6: An EMAGE screenshot: a typical set of gene expression information presented via the web interface (with header and footer removed).

The subsequent five sections impart the experimental results. The first gives the actual result of the experiment - a series of images illustrating where the gene is expressed. The images are of a slice of a particular mouse. The following **Notes for interpretation** presents a key for the annotations used in the images.

By examining the images, an expert is able to identify which structures[5] the gene is expressed in. The researchers do this for their journal publication by creating textual annotations. This analysis is reported in the segment called **Sites of Gene Expression Annotated Manually**. Three are provided, along with an indication of the level of expression, e.g. strong.

A manually produced mapping of the above images to a 3D model creates the **Spatial Annotation** section. This starts by displaying the actual spatial mapping performed. Subsequently, a list of quality measures is supplied. "Data pattern clarity and extraction" signals how clear the experimental image is. "Morphological match of data embryo to template model" indicates how well the subject in the image relates to the standard 3D model used for spatial annotations. Lastly it states who approved the spatial mapping: it can be either the researcher or the EMAGE editors.

Next comes **Sites of Gene Expression Inferred by the Spatial Mapping** - this is the list of textual annotations derived from the spatial annotations. A list of relevant structures is shown, as is the volume of the structure that has the gene expressed in it. This is shown as a percentage for each level of expression. For example, 1.5% of extraembryonic ectoderm has *fgf5* strongly expressed, and the remainder of the tissue (98.5%) does not contain the gene.

Authors lists the people responsible for the different information on the page. Firstly, the researcher is credited and their research paper cited (if one exists). "Indexed by" indicates who took the experimental result and mapped it to the EMAP anatomy ontology. It may be the researcher, the GXD editors, or the EMAGE editors (for more information on GXD see section 3.5.2). Finally, the reader learns who created the spatial mapping, normally this is an EMAGE editor, but it may be the researchers.

Submitted to EMAGE by specifies who submitted the experiment for inclusion

[5] Although there is disagreement in the biomedical world as to the exact definitions of the terms *tissue* and *structure* they will be used as synonyms in this document.

in EMAGE. EMAGE supplies tools that allow researchers to do this directly, and the EMAGE editors read published papers and submit data themselves. Historically, they have shared data with GXD, and verified it before it is included in the database. In this instance, the "Indexed by GXD" from the previous **Authors** section notifies a reader that GXD read the published paper, mapped the results to the EMAP anatomy, then shared the data with EMAGE. The "Submitted to EMAGE by . . . EMAGE Editor" indicates that the editorial team have reviewed and accepted the data from GXD.

The following two sections provide provenance information for the experiment. Ideally the **Probe** section imparts enough information for the same probe to be used by someone else. The information presented here will be taken from the article in which the experiment is published.

Specimen gives details of the experimental subject. The age of the mouse is given in two forms: Theiler Stage; and days post conception. Its "Genotype" indicates that this is the standard mouse with no genetic mutations.

The penultimate section presents a reference for the paper in which this experiment was published. In this case there are two publications, because the second one contains details of the probe used in the current experiment (the first citation).

Links presents connections (URLs) to related information in other resources. Notice, another link to the MGI (MGI:1930524). Clicking on this displays the page shown in Figure 3.7.

3.5.2 GXD

Mouse Genome Informatics (MGI) contributes a range of functions and online resources. One of these is a gene expression database called *GXD*.

Like EMAGE this resource has a web interface. It is fully integrated with the other MGI databases and it is not obvious where GXD stops and the other resources start. The complete suite delivers information on genes (linked to by EMAGE), proteins, probes, structures, gene expression information, pathways, protein functions, and a wide variety of links to other resources such as the Gene Ontology[6] - an ontology that provides a mechanism for describing genes.

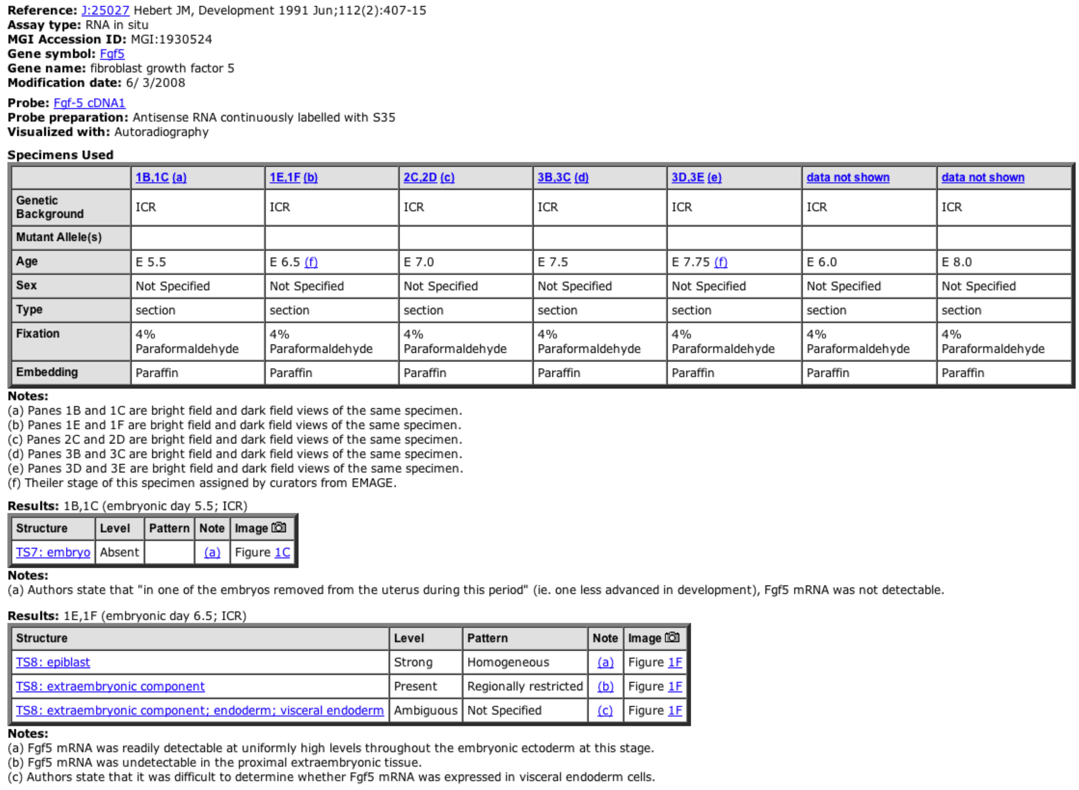
[6] `www.geneontology.org`

Behind GXD is the EMAP anatomy. In places the original anatomy has been extended to deliver finer granularity (lower level nodes). Furthermore, GXD has extended the ontology to cover the adult mouse (TS28).

Unlike EMAGE, GXD only produces textual annotations. It has a different organisation and presentation of information. To illustrate this, Figure 3.7 is the web page for the experiment in Figure 3.6 - due to the size of the actual web page, this is merely a subset of it. Immediately it is noticeable that the page contains details of more than one procedure. There are seven specimens used, and seven different procedures performed. Results are given for each procedure on this page - though most are not shown in Figure 3.7. EMAGE has the policy of splitting such procedures up, so that each one has a different EMAGE ID.

Another difference between EMAGE and GXD is the content in terms of experimental types. EMAGE only contains experiments that supply spatial information whereas GXD attempts to supply a more complete picture of gene expression and as such contains a far wider range of experiments including Northern blot, Western blot, RNAse, and RT-PCR.

Despite the obvious differences in presentation, and the lack of spatial annotations, both resources publish similar information, accordingly Figure 3.7 shall not be discussed further. For more information on GXD see Smith *et. al.* [132].



Reference: J:25027 Hebert JM, Development 1991 Jun;112(2):407-15
Assay type: RNA in situ
MGI Accession ID: MGI:1930524
Gene symbol: Fgf5
Gene name: fibroblast growth factor 5
Modification date: 6/ 3/2008

Probe: Fgf-5 cDNA1
Probe preparation: Antisense RNA continuously labelled with S35
Visualized with: Autoradiography

Specimens Used

	1B,1C (a)	1E,1F (b)	2C,2D (c)	3B,3C (d)	3D,3E (e)	data not shown	data not shown
Genetic Background	ICR	ICR	ICR	ICR	ICR	ICR	ICR
Mutant Allele(s)							
Age	E 5.5	E 6.5 (f)	E 7.0	E 7.5	E 7.75 (f)	E 6.0	E 8.0
Sex	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified	Not Specified
Type	section	section	section	section	section	section	section
Fixation	4% Paraformaldehyde	4% Paraformaldehyde	4% Paraformaldehyde	4% Paraformaldehyde	4% Paraformaldehyde	4% Paraformaldehyde	4% Paraformaldehyde
Embedding	Paraffin	Paraffin	Paraffin	Paraffin	Paraffin	Paraffin	Paraffin

Notes:
(a) Panes 1B and 1C are bright field and dark field views of the same specimen.
(b) Panes 1E and 1F are bright field and dark field views of the same specimen.
(c) Panes 2C and 2D are bright field and dark field views of the same specimen.
(d) Panes 3B and 3C are bright field and dark field views of the same specimen.
(e) Panes 3D and 3E are bright field and dark field views of the same specimen.
(f) Theiler stage of this specimen assigned by curators from EMAGE.

Results: 1B,1C (embryonic day 5.5; ICR)

Structure	Level	Pattern	Note	Image 📷
TS7: embryo	Absent		(a)	Figure 1C

Notes:
(a) Authors state that "in one of the embryos removed from the uterus during this period" (ie. one less advanced in development), Fgf5 mRNA was not detectable.

Results: 1E,1F (embryonic day 6.5; ICR)

Structure	Level	Pattern	Note	Image 📷
TS8: epiblast	Strong	Homogeneous	(a)	Figure 1F
TS8: extraembryonic component	Present	Regionally restricted	(b)	Figure 1F
TS8: extraembryonic component; endoderm; visceral endoderm	Ambiguous	Not Specified	(c)	Figure 1F

Notes:
(a) Fgf5 mRNA was readily detectable at uniformly high levels throughout the embryonic ectoderm at this stage.
(b) Fgf5 mRNA was undetectable in the proximal extraembryonic tissue.
(c) Authors state that it was difficult to determine whether Fgf5 mRNA was expressed in visceral endoderm cells.

Figure 3.7: A GXD screenshot: a typical set of gene expression information. This is part of the page for the experiment shown Figure 3.6.

3.5.3 Cancer Genome Anatomy Project

The *Cancer Genome Anatomy Project*[7] (CGAP) has the goal of determining the genes responsible for causing cancer in different types of cells. This resource contains information on RNA interference[8], biological pathways, genes, tissues, and SAGE experiments. Only the SAGE data is used in this thesis - for a brief description of the SAGE technique see Section 3.4.2.

The SAGE data is for mouse and human, and features both libraries with long tags and libraries with short tags. Much of the mouse data came from the *Mouse Atlas of Gene Expression Project*[9]. This resource focuses on providing a numerical count of the amount of a gene expressed in a particular area of the mouse. Associated

[7] `http://cgap.nci.nih.gov`

[8] A process that effects the rate at which a piece of RNA is translated into a protein. As such, it has a similar effect to gene regulation.

[9] `www.mouseatlas.org`

with each library is an image of the structure(s) from which the library was produced.

The anatomy used by CGAP is their own proprietary anatomy for the developmental mouse.

3.5.4 Allen Brain Atlas

The *Allen Brain Atlas*[10] (ABA) was set up to research gene expression in the mouse and human. Currently it focuses on the mouse's brain and spinal cord, and the human cortex - only publishing experimental results that the project has produced itself.

Concentrating on the mouse's brain, the project experiments on both the developmental[11] and adult mouse[12]. These are spatial experiments (*in situ* hybridisation with MRA levels), the results of which are associated with the ABA's proprietary anatomy and digital atlas.

Because ABA is funded privately by a philanthropist there is no requirement for all the ABA data to be published. Accordingly, the data made publicly available is a small subset of the total collected.

3.5.5 GENSAT

GENSAT[13] is another resource dedicated to the developmental and adult mouse central nervous systems (i.e. the brain). GENSAT uses *in situ* techniques to capture gene expression information, which the project then publishes in its own database. GENSAT is funded by U.S. government bodies, and thus the entire database is made publicly available.

3.5.6 Microarray resources

ArrayExpress[14] and *GEO*[15] are dedicated to publishing information from microarray experiments. These resources publish experimental results from a variety of or-

[10] `www.brain-map.org`

[11] `http://developingmouse.brain-map.org/`

[12] `http://mouse.brain-map.org/`

[13] `www.gensat.org`

[14] `www.ebi.ac.uk/microarray-as/ae/`

[15] `www.ncbi.nlm.nih.gov/geo/`

ganisms, including the developmental mouse. ArrayExpress and GEO share data, consequently their content is relatively similar.

As array-based experiments are not considered in this work these resources shall not be reviewed further.

3.6 Summary

Previously in this chapter the biological background necessary to understand this thesis is discussed. This document's use case is *in situ* hybridisation gene expression for the developmental mouse.

Gene expression information notifies a reader which genes are responsible for the development of particular anatomical features such as the mouse's tail. *In situ* gene expression experiments indicate exactly where a gene is expressed in the unborn mouse - ultimately bestowing an image of the exact location. Other types of gene expression experiment exist, and some of these are mentioned briefly.

Once performed, gene expression experiments may be published in a journal, and then their results re-published in a series of specialised online resources. A number of such resources are discussed, with particular emphasis given to those featured in this document: EMAGE and GXD.

Chapter 4

Online biological resources: problems and solutions

The chapter shall focus on the motivation for this thesis, firstly describing the problem and then outlining a possible solution. Work concentrates on one particular area of biology: *in situ* gene expression for the developmental mouse; however, it is applicable to biology in general. The following quote from Antezana *et. al.* relates to the field of biology:

> In the post-genomic era, life sciences (LS) turned into a very data-intensive domain. Therefore, scientists in this domain are facing the same challenges as in many other disciplines dealing with highly distributed, heterogeneous and voluminous data sources. However, these problems are more acute in the LS compared to many other domains due to a number of factors. Those factors could be divided into two groups. One group that could be called natural, reflects the specifics of biological data and knowledge, first of all its complexity [133]. The other group of factors could be categorized as cultural. The avalanche of data outpouring from the high-throughput genome-wide technologies caught life scientists unprepared. . . . The result was an abundantly idiosyncratic domain specialization. In particular, the biomedical domain is plagued with extremely high fragmentation. ([134] page 395)

Although the quote clearly describes the wider biological domain, the content of this chapter will illustrate that the description is equally valid for the use case described in Chapter 3.

The interconnectedness of biology ensures that the field of gene expression is just as intricate as any other. This complexity is commonly realised as inconsistency. Conflict occurs in a number of situations, for instance, between experiments that seemingly should produce the same result, but in reality do not. Section 4.1 considers this from the perspective of the use case.

Furthermore, the sheer scale of the biological world means that our knowledge of it is incomplete, with a significant number of gaps requiring investigation. Again, the situation is identical within the use case - as discussed in Section 4.2.

In Section 3.5 a range of resources providing gene expression information for the mouse are reviewed: EMAGE, GXD, GENSAT, ABA, GEO, and ArrayExpress. Most of these resources use their own anatomies, and their own online publishing mechanism (i.e. a proprietary database schema, web site, and programmatic interface). In short, fragmentation is just as serious an issue for the use case as Antezana *et. al.* state it is for the wider biological realm. Section 4.3 shall probe the distributed nature of the use case.

Subsequently, Section 4.4 outlines the high level objective of this work. In Section 4.5 the reasons behind the use of argumentation shall be explored. Penultimately Section 4.6 discusses some of the issues to be investigated in this work, and the chapter is concluded with a summary in Section 4.7.

4.1 Inconsistent data

When two identical experiments are performed, instinctively it is expected that the outcomes will be identical. Yet, the actual results may contradict one another - experiment one may suggest that a gene is expressed in a tissue, while experiment two indicates that the gene is not expressed.

Such contradictions can occur for a variety of reasons. Firstly, the complexity of the experiments means that many parameters can affect the outcome. Ensuring that two experiments are truly identical is nigh on impossible when something as apparently minor as a brief 1°C temperature change can alter the result[1]. Secondly, experiments are performed and analysed by humans introducing the possibility of human error, such as incorrectly identifying the age (developmental stage) of the mouse. Thirdly, the outcome of an experiment is evaluated and mapped by a human creating the prospect of differences of opinion. For example, when the result of an experiment is as vague as Figure 4.1 it is understandable that two people may disagree about the conclusion.

When two supposedly identical experiments seem to disagree, it is easy to say that they are not actually identical and thus there is no conflict. Yet this is not the attitude of the biologists. They perceive such a disagreement as an inconsistency that needs to be resolved. Accordingly the same viewpoint shall be adopted during this work.

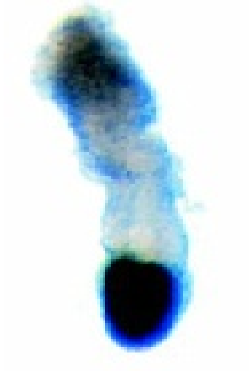


Figure 4.1: A result from EMAGE experiment EMAGE:821.

Regrettably inconsistencies are common in biology, which creates a problem for resources such as GXD that republish experimental results. Determining which result is correct would be problematic as it would involve the use of personal opinion and judgements that may not be shared by other biologists (alternatively the experiment could be repeated, but the resources are not equipped to do so). Consequently, resources such as GXD publish all the experiments that meet their quality requirements. This effectively places the onus for the resolution of contradictions onto the biologists who use the resource.

[1] This is related to the amount of energy required to break the pre-existing chemical bonds so that the probe can bond with its target RNA/protein. More heat means more energy is available, thereby directly increasing the ability of the target and probe to bond.

In some cases the solution is simple. For example, GXD contains one annotation suggesting that the gene *Tnc* is not expressed in the mouse brain in TS24 and fourteen suggesting *Tnc* is expressed.

At other times the settlement is far more complex. Consider the gene *Bmp4* in the eye TS22. At the time of writing, GXD features one annotation[2] suggesting the gene is expressed, and one annotation[3] suggesting it is not expressed. Tackling this disparity requires a closer examination of the data in GXD, and possibly a perusal of the published papers.

An examination of GXD in 2007 suggested that the resource contained some 1300 inconsistent annotations [135] - it must be stressed that this is not an issue unique to GXD, it is used merely to illuminate the discussion. GXD continues to publish new annotations leading to a gradual increase in the number of inconsistencies.

In summary, it is essential to query for genes expressed and not expressed in the tissue of interest and manually resolve the resulting inconsistencies. However, as Sections 4.2 and 4.3 show, it is necessary to repeat this for multiple resources.

4.2 Incomplete data

Many biological resources, including those in the field of *in situ* gene expression for the developmental mouse, are incomplete because they do not publish an entire view of their chosen domain.

To be a complete resource, it would require to be more than an *in situ* hybridisation resource such as EMAGE, it would entail covering the full range of gene expression experiments. Furthermore the resource would need to publish details from every single gene expression experiment undertaken on the mouse. In addition, these experiments must, collectively, have used every technique to look for every gene in all mouse tissues. This is a colossal volume of data, and it is not surprising that no resource accomplishes this.

Some might argue that any resource that does contain all the above information is still deficient as it does not consider mutants. For them, a complete resource would

[2] GXD accession ID = MGI:3039849.

[3] Accession ID = MGI:3756832.

contain the above information for the wild type mouse, plus the same information for every single mutant mouse.

Consider three resources in the use case domain: GXD, EMAGE, and ABA. The final resource is incomplete because it specialises in one class of experiment on only one part of the mouse (the brain). Therefore if a user wishes to know which genes are expressed in the leg it cannot help them. ABA was set up by a project to publish the experimental results of that project. Currently, the resource publishes only a fraction of all the data collected by the project - creating another respect in which this resource is incomplete.

EMAGE is incomplete because it publishes only experiments examining the location of the expression, thus ignoring experiments focusing on the quantity of expression. It only considers the developmental mouse, and thus is missing information because it does not consider the adult mouse.

GXD publishes a wider range of experimental results, and covers all 28 Theiler Stages of the mouse - it provides a richer, fuller, picture of the domain. However, it too has gaps in its coverage. GXD publishes experimental results from the community at large - as opposed to ABA which is a proprietary database publishing results for one project. Therefore, for GXD to be complete it would need to publish every single gene expression result for the mouse. However, this is not the case. Often large scale projects, such as ABA, only publish their data in their own resource(s). In addition some experiments will not be submitted to the GXD editorial team, and will "slip under the radar". Furthermore, some experiments will be rejected because the editorial team do not believe that they satisfy the quality metrics desired by their resource.

In summary, it is impossible to obtain a complete picture of gene expression in all mice (wild type and mutants). Moreover, to gain as comprehensive a picture as possible, it is necessary to consult a wide range of resources.

4.3 Distributed data

A third problem associated with biological data is its distributed nature. This is something that has already been broached in Section 4.2 where EMAGE contains one set of data and GXD a second set.

EMAGE and GXD contain significant overlap because both resources publish experimental results from third parties - this is depicted by Figure 4.2. Notice that Figure 4.2 illustrates that the resources overlap in one particular area (*in situ* experimental results), and that the overlap is not absolute. Consequently, some spatial experimental results are found in GXD but not EMAGE and vice versa. Thus a biologist wishing to gain an understanding of the domain must consult both resources.

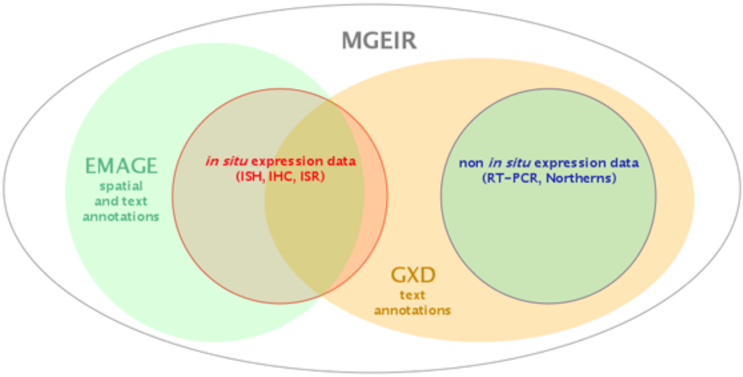

MGEIR
EMAGE
spatial and text annotations
in situ expression data (ISH, IHC, ISR)
non *in situ* expression data (RT-PCR, Northerns)
GXD
text annotations

Figure 4.2: Illustration of the overlap between EMAGE and GXD from `www.emouseatlas.org/emage/about/mgeir.html`.

In addition to creating inconsistencies within a single resource, the issues described in Sections 4.1 and 4.2 may cause inconsistencies between different resources. For example, when asking if the gene *Otx2* is expressed in the primitive streak in TS9, at the time of writing EMAGE has one annotation[4] providing the answer "yes". Yet GXD has one annotation[5] advancing the answer "no". Hence it is very difficult to determine whether or not *Otx2* is expressed. The situation is confusing because the resources are citing the same experiment, i.e. the resources have analysed the experimental result (shown in Figure 4.1) in different ways. This highlights the level of subjectivity involved in assessing a result.

Such conflict reinforces the message that neither resource can be used on their own, and that care has to be taken when posing a query. Simply asking which genes are expressed in primitive streak TS9 will provide only the EMAGE opinion. Unless the opposite query is posed too, the conflict will not come to light.

[4] Accession ID EMAGE:821

[5] Accession ID MGI:2154979

4.4 Objectives

Sections 4.1 to 4.3 illustrate the problems faced by a biologist when trying to determine if a gene is (not) expressed in a particular tissue of the developmental mouse.

The ambition of this work is to improve the understanding of the applicability of computational argumentation within biology through the study of a particular biological use case involving *in situ* gene expression for the developmental mouse. Whilst tackling the problems faced in the use case, insights into the future role of computational argumentation in biology shall be sought.

4.5 Why argumentation?

The previous section clearly states that the objective of this work includes the use of argumentation. However, it did not state the reasons for using argumentation - they shall be outlined here.

4.5.1 Requirements

Before commencing the process of solving the issues previously discussed in this chapter, it is worthwhile contemplating what would constitute a solution.

To start with, it is assumed that an application will be created to provide the biologists with the support they need. Yet what task should this system perform?

It is clear that incomplete and inconsistent information needs to be dealt with. Additionally, it seems natural to consider that the goal of the system should be to make a decision. Yet this is not the case. Decisions regarding gene expression are controversial, therefore it would be necessary to customise a system to each user, in order to make a “correct” decision. However, this would not be enough, as ultimately it does not indicate *why* the decision is made. Moreover, in all likelihood a final resolution will come with further biological research. Any solution attempted here would be little more than informed speculation based on a partial view of the world. As such the aim should not be to provide a definitive conclusion, but instead to consider what inference seems most likely with the current state of the domain. As new experiments are performed, new annotations will become available, and thus the inference will need to be reconsidered.

To be of genuine use to the biological community a system must explode the reasoning process providing unobscured access to the underlying information and knowledge on which the inferences are based. This allows the user to treat the system as a platform on which to base their own thought processes. Supplying direct links to the underlying data (sources), will strengthen the platform constructed, by allowing the user to conduct any further research that they may desire.

Finally, in order to be trusted, any system must use techniques that are clear to the user group - many of which do not have the time, appetite or the necessary background to comprehend complex mathematical or logical solutions.

4.5.2 Alternatives

Manifestly a wide range of technologies (and thus research) is applicable to the above problem. Although decision making has previously been rejected, instinctively there appears to be a need to make and then explain a decision, thus decision making and explanation mechanisms may be appropriate. As data is being brought together from a variety of heterogenous sources, data integration technology must be considered too.

Data integration

Data integration is a research domain interested in:

> In general, integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system. ([136] page 39)

Outwardly this is very similar to the aim of this work. However, there is one clear difference: the goal of this work is not to create a "unified new whole". Instead of unifying the data, the idea is to create a unified inference based on that data, i.e. whether or not the gene is expressed. Yet, in this work the unification is less important than the reasons for the unification. Furthermore, the idea of combining all the data into a "new whole" is unlikely to be popular with the data sources who provide the information. Those sources spend valuable human and financial resources gathering and annotating data. Naturally they expect, and deserve, credit for their work.

One domain related to data integration investigates the resolution of conflicts in data sources. Clearly, in that area, there are parallels with this work. Unfortunately, most so-called *inconsistency management* solutions suffer from one fundamental problem:

> . . . almost all past approaches proceeded under the assumption that there was some epistemically correct way of resolving inconsistencies or reasoning in the presence of inconsistency. More recently, [137] argued that inconsistency can often be resolved in different ways based on what the user wants ([138] page 367)

As discussed in Section 4.5.1, deciding whether or not a gene is expressed is often controversial. Thus the traditional answers from this domain are clearly unsuitable. However, there are solutions that attempt to rectify this problem. For example, Martinez *et. al.* [138] suggest creating a number of so-called policies which allow the user to select their own resolution. Martinez and Hunter [139] further this idea by using computational argumentation to decide which policy is most appropriate for a particular context.

At first glance Martinez's work seems to provide an attractive panacea to the issues discussed in this document. However issues remain. Firstly, she is dealing only with inconsistent information. Secondly, her basic idea is that a finite number of functions can be developed that allow all inconsistencies to be resolved - when considering human opinion in a complex domain this is questionable. Finally, the principle is still very much in the research phase and is not yet ready for real world deployment. This means that there is no solution ready for trial to ascertain whether or not a biologist can understand how and why these policies work for them.

Decision making

Like all the domains discussed in this section, decision making is a large area of research with numerous proposals and solutions. Properly documenting this field is beyond the scope of this work, and a brief profile is offered instead.

Decision making techniques can be split, broadly speaking, into four categories:

certainty deterministic knowledge;

risk complete probabilistic knowledge;

uncertainty partial probabilistic knowledge;

ignorance no probabilistic knowledge.

In this use case there are no known, measured, and verified probabilities. Yet, it is possible to approximate such knowledge by consulting an expert. Although these expert assignments are controversial, they may act as a basis for conducting decision making under uncertainty. Alternatively, it may be possible to learn the probabilities from a training set. Either of these approaches would allow a range of technologies to be adopted, including the popular Bayesian nets [140].

Intuitively there is no reason why Bayesian networks could not be applied to the current problem. Indeed Friedman *et. al.* [141] use this technology to investigate gene expression for microarray data (see Section 3.4.2). Their aims are more sophisticated than those proposed for the current research:

> Our aim is to model the system as a joint distribution over a collection of random variables that describe system states. If we had such a model, we could answer a wide range of queries about the system. For example, does the expression level of a particular gene depend on the experimental condition? Is this dependence direct or indirect? If it is indirect, which genes mediate the dependency? ([141] page 607)

The work of Friedman *et. al.* proves that Bayesian nets can be used with gene expression data. However, that does not mean their work may be applied to the current task. In order to construct their model of gene expression they use the following variables:

> In this paper, we are dealing mainly with random variables that denote the mRNA expression level of specific genes. However, we can also consider other random variables that denote other aspects of the system state, such as the phase of the system in the the cell-cycle process. Other examples include measurements of experimental conditions, temporal indicators

> (i.e., the time/stage that the sample was taken from), background variables (e.g., which clinical procedure was used to get a biopsy sample), and exogenous cellular conditions. ([141] page 607)

They are able to obtain all this information because they are using results from microarray experiments. Traditionally, these experiments are well documented, with adopted guidelines[6] defining the minimum provenance information to be captured and associated with the experiment's publication in both journals and online data sources, e.g. ArrayExpress.

This work considers *in situ* hybridisation (ISH) experimental results. There is a standard[7] for documenting this class of experiment too - disappointingly, the research community does not strictly adhere to it. Consequently, far less provenance information is collected in comparison to array-based experiments; examining Figure 3.6 in Section 3.5.1 it is obvious that scant information regarding the experimental technique and sample is provided. Therefore a number of the variables used by Friedman cannot be applied.

The results of ISH and array-based experiments differ further in one crucial way. The main variable used by Friedman was the mRNA expression level of a specific gene. In array-based experiments this is quantified precisely as a number; however, ISH experiments only provide a vague natural language description such as "moderate" or the even less descriptive "present".

With far less data available on which to base a model, it is questionable if the Bayesian approach adopted by Friedman *et. al.* would be applicable to the current use case.

Another relevant work is that of Jeffreys *et. al.* [9]. Jeffreys *et. al.* suggest that there are two tasks in the use case they are considering: analysing the results of a protein structure prediction tool. Firstly, something must generate the predictions; secondly, the predictions are analysed by a human expert to provide a final decision on their validity. Their work is aimed squarely at the latter task, in effect replacing the human expert. When comparing Bayesian nets to argumentation for this task the authors remark:

[6] `www.mged.org/minseqe/`

[7] `http://scgap.systemsbiology.net/standards/misfishie/`

> Our aim is to model the reasoning process of the researcher who uses the tool and evaluates its output. Whilst human reasoning of this kind accommodates a degree of uncertainty, it does not use a complex probabilistic model. Bayesian nets and argumentation frameworks are thus not direct competitors. Argumentation theories have been developed as an approach to reasoning about uncertainty where even rough estimates of probabilities are not available or meaningful. ([9] page 932)

Parallels can be drawn betweens Jeffreys *et. al.* [9] and the current work. Although the data sources, information, and tasks are different, in both cases the idea is to replace, or enhance, analysis performed by a human expert. Consequently, the reasons why Bayesian nets are not appropriate in Jeffreys *et. al.* [9] renders them, and other probability-based decision making techniques, redundant for this work.

Explanation systems

Closely related to the notion of making a decision, is the idea of explaining that decision to the user. As an example consider the work of Williams and Williamson [81] in which argumentation is used to generate explanations for decisions made with Bayesian nets.

Irandoust summarises the main purposes of explanations:

> . . . one could say that explanations can be used to: (i) assure the user (or ultimately convince her) that the systems reasoning is logical and its conclusions sound, relevant and useful; (ii) provide visibility into the systems states, actions and intentions and guide the user in performing her problem-solving tasks (iii) teach or give the user the possibility to learn by exposing the systems domain knowledge and reasoning techniques. ([142])

Many argumentation scholars, such as Walton [143], believe that the domain of argumentation can help AI understand and better implement explanation-aware systems. The two concepts are remarkably similar, with Walton providing the following distinction:

> ... the purpose of an argument is to give a reason to support a claim made by one party in a dialogue. The claim is something that is doubted by the

> respondent in the dialogue. It is a proposition that is at issue or unsettled. An argument is supposed to present a good reason for the respondent to come to accept this proposition as true, thus removing the doubt. The purpose of an explanation is to help the questioner who doesn't understand something. ([14] page 76)

Research in argumentation and explanation has overlapped for a considerable period of time - in 1992 Wick [144] noted that explanation research had taken the Toulmin scheme for their own:

> It is interesting to reflect on how research in expert system explanation has, without stated intent, evolved to consider each piece of knowledge in an expert system as a single Toulmin structure. (page 167)

The high level of similarity means that argumentation structures and techniques are used often for generating explanations:

> . . . a number of works have advocated use of argument structures as a basis for knowledge representation [145]; and in particular characterisation of computer generated explanations as arguments, [146, 147, 148]. Indeed Ye and Johnson [149] describe empirical studies demonstrating that argument structured explanations are the most effective in terms of bringing about changes in users attitudes toward rule-based systems. ([150] page 33)

The use of argumentation in explanation-aware systems is particularly beneficial when the user is sceptical of the advice given by the system and requires some persuasion: an example being the medical systems that try to explain why it makes sense to adopt a healthier lifestyle, e.g. [11, 108].

In summary, argumentation and explanation frequently use the same technology. This is possible because the two concepts are extremely similar, the difference being contextual - arguments seek to persuade whereas explanations aim to elucidate. Consequently it seems that if an explanation is necessary, one of the best ways of providing it is to exploit argumentation.

4.5.3 Motivation for argumentation

The current use case requires the ability to reason with inconsistent and incomplete information. Inferences drawn from this information are controversial, and need to be revised as new information becomes available. In essence, a presumptive, plausible, and defeasible form of reasoning is required. Non-monotonic reasoning solutions are designed for such a situation. As Section 2.2 reports, argumentation is one of a number of ways of performing non-monotonic reasoning. Dung highlights the similarities of these different forms:

> ... many of the major approaches to nonmonotonic reasoning in AI and logic programming are different forms of argumentation. ([31] page 4)

One of the main rationales for choosing argumentation over other styles of reasoning is the intuitiveness of the basic concept. Although the theory behind argumentation may be complex and rely on logic, the basic concept is an everyday human practice. Everyone can understand the idea of people debating a topic, and it is not difficult for them to imagine a computer doing the same. This simplicity confers an advantage over many other forms of reasoning, as it allows the system to be explained quickly and effectively to everyone, regardless of their background.

A further justification for this exploration of argumentation in biology, is the popularity of the practice within the medical informatics community. Inherently the two domains are similar - biology provides much of the scientific basis of modern medicine. Thus if argumentation can be used in medicine (see Section 2.6.2 for evidence of this) then it must be useful for biology too? The answer, of course, is ‘perhaps’. There is no doubt the two domains are interrelated; however, they are clearly different - a degree in medicine does not qualify a person as a biologist and vice versa. Therefore it may be that this difference effects the applicability of argumentation.

Finally Occam’s razor[8] contributes a reason to employ argumentation - this is the idea that simplicity is good. It is possible to use one technology to make an inference, then a second to explain that inference, and possibly a third to make a decision based on the inference. Alternatively, argumentation can be used to fulfil every role.

[8] www.merriam-webster.com/dictionary/occam’s%20razor

4.6 Remaining issues to be resolved

For argumentation to be applied successfully to biology a number of questions must be raised and resolved. It is the aim of this work to explore possible answers to these questions through the development of a solution to the use case issues described in Sections 4.1 to 4.3. This section will document the questions to be investigated. The detailed questions are given below, but first a summary is presented using the following abstract queries:

1. Which form of argumentation is appropriate as a solution to the issues described in Sections 4.1 to 4.3 - and how effective is it?
2. How should argumentation be presented to a biologist so that (s)he can understand, and utilise it?
3. What insights have been gained from the work, and how does that inform the future use of argumentation within biology?

The detailed questions are listed and explained in the following text.

The first abstract query asked what form of argumentation is appropriate, and how effective is it? Yet, it failed to consider the possibility that argumentation may not be appropriate at all. Might it be the case that simply generating reasons (arguments) for and against each side of a query will be sufficient without the added debate (argumentation)?

Alternatively, this first abstract query can be interpreted as asking what argumentation actually is? How shall it be implemented in this domain? What form can the argumentation realistically take? What practical constraints exist? What are the consequences of these constraints?

A number of implementation related questions may be posed under the banner of abstract query 1. For example, which architecture appears appropriate for this situation? In addition, how can this argumentation system be deployed in order that it is used by a range of biologists?

Abstract query 2 relates both to the presentation of arguments and argumentation. Traditionally argumentation theory presents arguments using a graphical representation. Several main forms of graphic presentation are featured commonly in tools and publications. Which of these is favoured by the biologists? Furthermore, is the graphical form actually preferable to a textual depiction? Moreover, how well do the standard argumentation concepts translate to the biological domain?

The final abstract query considers issues relating to the broader use of argumentation within biology - effectively what has been learned during this work? Assuming that this process represents the first step towards the widespread deployment of argumentation in biology, what foundations need to be laid before the common uptake of argumentation can take place?

Chapters 5 to 8 document the work undertaken to resolve the above questions, and may contain implicit answers to the above questions; however, explicit answers will not be provided until the work is analysed in Chapters 9, 10 and 11.

4.7 Summary

This chapter examines the problems of conflicting, incomplete and distributed information that affects two biological databases EMAGE and GXD. Although these databases are situated within the field of gene expression for the developmental mouse, the nature of biology ensures that these issues affect most domains to a degree.

Following on from this, the objectives of this work are stated, before a justification of the application of argumentation is provided. Finally a number of questions to be answered by this work are documented.

Chapter 5

Argumentation in biology - a conceptual consideration

Argumentation is a multifaceted research discipline that can be implemented in a variety of ways for a diverse range of purposes. Central to all of these approaches is the idea of an *argument* being fundamental - as a unit of reasoning, communication or modelling.

Biology is a productive discipline that generates a large volume of data each year. Chapter 4 explains why some of this information is inconsistent - furthermore it documents the reasons behind the incompleteness of resources that republish biological information online.

This document focuses on a use case in the world of *in situ* hybridisation gene expression for the developmental mouse. That domain epitomises the problems of inconsistency and incompleteness commonly found in biology. In the previous chapter the objective of this thesis was summarised as:

> ...to improve the understanding of the applicability of computational argumentation within biology through the study of a particular biological use case involving *in situ* gene expression for the developmental mouse. Whilst tackling the problems faced in the use case, insights into the future role of computational argumentation in biology shall be sought. (Section 4.4)

Any solution for the use case must align the terminology and practices of the biological world with the domain of argumentation. To do so it is necessary to contemplate the following:

- What is an argument? How are these arguments to be used?
- Where do arguments come from?
- What can be argued over? How can arguments be attacked?

This chapter shall deliberate on the questions posed here. The first section (Section 5.1) reconsiders the problem examining how it stipulates the abstract notion of the solution by defining what an argument is in this context. This theme is continued into Section 5.2 where the impression of an *argument* is explored further alongside the strategy for the argumentation process. Section 5.3 takes a pragmatic stance, considering the data available for use in argumentation, and in what manner it may be applied. Throughout this scrutiny, the notion of conflict is pursued. An example of argumentation in biology occupies Section 5.4. The theme of the chapter is concluded, in Section 5.5, with an outline of the solution and the rest of this document, before a summary of the chapter in Section 5.6.

5.1 What is an argument?

As discussed previously, the notion of an *argument* has an array of effectuations, but which are of most relevance in the current situation? To arrive at an appropriate conclusion it is necessary to dwell on what is being attempted.

Resources (re)publish experimental results - in this case *in situ* gene expression results. Commonly users search the resources by the experiment's conclusion, i.e. roughly 90% of EMAGE's users query the resource by asking where a particular gene is expressed[1]. Alternatively, via EMAGE's web interface, a user can ask:

- which genes are (not) expressed in a specific tissue;
- which genes are (not) expressed in an area of the mouse;
- which genes are associated with a GO term, and where are these genes (not) expressed?

[1] Figure obtained using Google Analytics in Autumn 2009.

GXD provides similar functionality, albeit presented differently.

When interrogating either of the use case's resources the user may come across inconsistent and/or incomplete data. Such phenomena are at best a time consuming annoyance and at worst a full stop to the user's research. By loosely aggregating the data from the resources some of the incompleteness may be alleviated. Unfortunately, this comes with the disadvantage of highlighting the inconsistencies between the resources.

With inconsistencies between the resources, and between experiments in the same resource, it is not enough to aggregate data. There must be some mechanism by which the inconsistency is tackled too. Magically pulling an answer from a black box is unlikely to result in a user trusting the solution: indicating a transparent mechanism is vital.

Manifestly there is a requirement to reason with the biological data, and thus work around its limitations. Equally clear is the need to communicate this reasoning to the user. In conclusion the following two notions of an *argument* seem appropriate:

- a natural language justification;
- an element implemented in a computer program for problem solving (reasoning) purposes.

The notion of an *argument* can be further specified by reconsidering the resources and how they are used. The queries listed previously in this section can be reduced to two forms:

1. Where is gene G (not) expressed?
2. Which genes are (not) expressed in tissue T in Theiler Stage S?

Accordingly, the argumentation must centre on these questions too. Regardless of which question is asked, the outline of the answer is the same: G is (not) expressed in T at S. With this perspective, arguments can be viewed as reasons why G is (not) expressed in T at S.

5.2 Where do the arguments come from?

Previously, the current chapter has defined an *argument* for the context of this work. This was achieved by considering the broad theme of the thesis, and refining the ideas within it until only a limited range of argument conclusions were considered:

- Gene G is expressed in tissue T at stage S;
- Gene G is not expressed in tissue T at stage S.

The arguments are obliged to fulfil two roles: one designed with the computer as the intended audience, and one for the biological end-user. These roles are not distinct. The ideas communicated to the user must be related to the ideas generated during the reasoning process. As such, the latter role is a ‘natural’ presentation of the former, and shall be derived from it - the true origin of the end-user’s arguments is the genesis of the computer’s arguments.

Arguments utilised by the computer must be constructed from the available information, in essence the data available in resources such as EMAGE and GXD - this shall be analysed in Section 5.3. However, to generate arguments it is necessary to have raw information and knowledge of the correct way to analyse that information. Often such knowledge is not documented and instead exists as the intellectual property of domain experts. A summary of this discussion is presented in Figure 5.1. The summary is simplified, as before arguments can be constructed for the argumentation process, the knowledge must be extracted and modelled in a suitable representation for the argumentation system.

From this viewpoint, an argument is a defeasible inference because the expert’s knowledge, and the domain information, are both defeasible. The expert will extend, and refine his/her knowledge over time affecting the reliability of inferences based on previous knowledge. Likewise, the domain information will change. Furthermore, in both cases the possibility of errors cannot be ruled out - creating the hazard of erroneous arguments.

Conceivably this is not the only instantiation of argumentation that may work. Rather than using gene expression resources (database and human) exclusively, it might be possible to generate arguments that consider the expression level of a gene

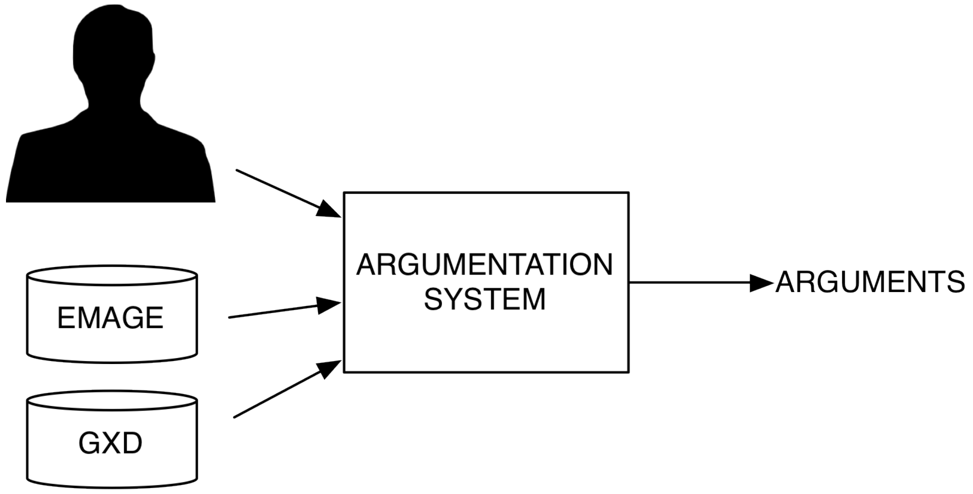


Figure 5.1: A high level examination of the solution proposed in this work. Expert knowledge is combined with biological facts, and wielded to generate arguments.

based on the underlying biological processes, rather than an analysis of the outcome of those processes. In this circumstance, the biological processes of translation and transcription would need to be modelled, as would all the subprocesses such as alternative splicing[2]. The sheer complexity of this is overwhelming. It is likely that many different argumentation processes would need to be undertaken and aggregated into a debate on whether or not the gene is expressed. For example, one debate may consider if the DNA sequence was accurate, a second may focus on the transcription process, a third on the splicing of that RNA and so on. In addition to requiring considerably more resources, this approach has the same basic flaw as the technique illustrated in Figure 5.1.

In order to debate any of the biological processes, knowledge of that process and the method necessary to analyse it is required. Regardless of whether that knowledge is acquired from a textbook or a direct face-to-face meeting, it emanates from a human being. Accordingly, there is the question of how reliable that person is. Even if this person is a world renowned expert, gaps in existing knowledge ensure that no person, or persons, could generate a perfect model of every biological process. One solution would be to argue over the reliability of the expert and when their knowledge can be applied; however, the model of expert reliability might itself be open to question, and thus subject to an argumentation process. Clearly a line must be drawn beyond

[2] A single RNA can be translated into several different proteins because of alternative splicing.

which weaknesses are acknowledged but not rectified.

The approach of arguing about the low level mechanics of gene expression demands far more work than can be achieved within one project. Accordingly, this work shall proceed by using expert knowledge of how to evaluate the data in EMAGE and GXD.

5.3 What can be argued over? *Experiment versus annotation*

Previously, in Section 5.2, the strategy for the argumentation process was chosen. The discussion continues by considering the data available for arguing and how it may be applied in the generation of arguments and counterarguments. The following terminology, although not widely used in biology, shall be used in this work to aid clarity:

Experiment The actual research undertaken including the experimental result (an image for *in situ* data);

Result The published outcome, i.e. the researcher's interpretation of the experimental result;

Annotation The (EMAGE or GXD) editors' analysis of the experiment, including their own interpretation of the result - may be textual or spatial;

Verdict The statement that a gene is (not) expressed in a tissue, which is either stated directly in an annotation or may be derived from an annotation. The term may be used to refer to the final conclusion drawn by the user.

The relationship between the concepts is documented in Figure 5.2. In an abstract sense, an experiment is performed and analysed to produce a result. The combination of the experiment and the result provides the source material for the annotation. That annotation contains limited provenance information and an analysis of the result. This analysis will contain one or more conclusions, e.g. gene *bmp4* is expressed in the future brain, and *bmp4* is not expressed in the neural fold. This information will lead to a user inferring a particular verdict, e.g. *bmp4* is expressed in the future brain.

Information on the raw experiment is stored in the journal article (assuming it was not part of a large screening program), and as such is not available unless the journal article can be textmined. Consequently, this work will access the experimental information contained within the annotation, as illustrated in Figure 5.2 through the use of a double line defining the extent of the system.

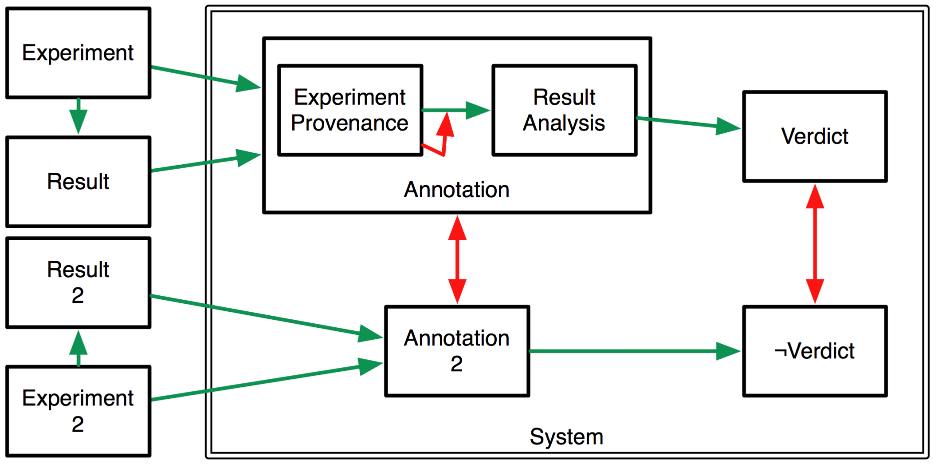


Figure 5.2: Illustrating the relationships between the original experiment, the original (published) result and the notions of annotation and verdict used in this work; green lines indicate a support or inference relationship with red lines depicting attack.

When viewed informally from an argumentation stance, the experiment and the result combine to allow the annotation to be generated. Thus if the experiment and result are trustworthy it increases the likelihood that the annotation is correct. In a similar vein, the annotation directly supports the inference of the verdict. These relationships are demonstrated in Figure 5.2 by the use of green lines. Considering the annotation in more depth: the analysis of the result is supported by the provenance information. The notion of *experimental provenance* contains all the data the resource holds on the experiment including the researcher, the probe, the sample, and the genuine result (image) produced by the experiment.

Normally the annotation is a subset of the information in the experiment and result, possibly with a small extension provided by the resource. However, it is possible that one experiment, when published, refers to a second experiment, in which case the annotation may be generated from two experiments and results. Occasionally, when examining the experimental result, the database's editors decide they disagree with the published result. As the experimental result is part of the experiment, the resulting annotation can still be considered a subset of the information in the experiment-result pair.

There is a difference between the two relationships: experiment/result to annotation; and, annotation to verdict. The annotation is, in some form, a view (in the database sense) over the experiment and result. Therefore the experiment is not a reason to believe the annotation as they are, at some level, equivalent; however, reasons to trust the experiment are automatically reasons to trust the annotation. In contrast, the annotation does provide a reason to believe the verdict. This occurs because the verdict is based either on the raw provenance information[3] (i.e. the image taken as the result of the experiment) or on the resources' analysis of the result. Either way, the verdict is inferred from the provenance information rather than simply being a re-publication of raw information.

There is another way of examining this difference in the relationships. Without an experiment there is no annotation. Yet, a verdict - *bmp4* is expressed in the future brain of stage 12 - can exist, albeit irrational, without an annotation to support it. However, if the annotation does exist it is possible to use it as a justification of why the verdict is correct. As such, the latter is the traditional argumentation notion of support where one entity allows the other to be inferred with some degree of confidence. A more detailed consideration of the relationship between the experiment/result and annotation is tangential, as the raw experimental information is not available, and shall not be pursued.

Figure 5.2 uses red lines to indicate potential lines of attack within the area of the system[4], with a double headed line pointing at the potential for rebuttal and a single headed arrow indicating an undercut. Immediately, a red line seems to be missing: the line attacking the support between the annotation and the verdict. This is not included because such an attack is not viable within the current work.

To attack the support between an annotation and a verdict would imply that there was something wrong with the notion that "an annotation provides a reason to believe a verdict". Here the discussion is not about a particular annotation but the

[3] Experts often prefer to perform their own results analysis.

[4] Lines of attack outside the system are ignored.

principle in general. In biology the basic principle stands and is not deemed open to negotiation, thus such an attack is not considered here.

Considering the attack relationships in Figure 5.2, the most obvious is the rebuttal between the annotations. As annotations directly lead to verdicts, it is possible for the verdicts to contradict one another.

The final line of attack is between the experimental provenance and the result analysis. Without access to the underlying article (assuming there is one), the experimental provenance contains all the information that can be used to evaluate the experiment and the resource's analysis of that experiment. Clearly this information has the potential to influence the reader's confidence in the result analysis. Accordingly, there is both a green line of support and an undercut of that support depicted in Figure 5.2 between the notions of experimental provenance and result analysis. It is worth noting that if an annotation is defeated in this manner, and the verdict is no longer supported, that does not provide a reason to support the opposite verdict.

For instance, if a verdict suggesting that *bmp4* is expressed in the future brain, is defeated by removing the supporting annotation then at most it is possible to say that it is unknown if *bmp4* is expressed in the future brain. In contrast, if the verdict's only supporting argument is defeated by a stronger rebutting argument, then with the currently available knowledge, it is possible to claim that *bmp4* is not expressed.

Ultimately, an annotation is deemed to be a reason to support a verdict. The information contained within the annotation provides the reason to trust that annotation, and thus the verdict it supports. However, it may contain the information necessary to reject the annotation and its associated verdict too.

5.4 Arguing about *in situ* gene expression data - an example

In Chapter 6 the real arguments used in this work shall be discussed. Before that a straightforward example will illustrate the basic principles of arguing over the current use case - this will be a simplification of the real process.

During this example a PROLOG-like logic will be assumed - PROLOG [151] is a logic programming language often used for academic purposes. It has a simple protocol in which uppercase letters indicate variables, and lowercase letters are constants. Thus the constant bmp4 may be stored in the variable Gene.

emageExpressed(bmp4, hindlimb) indicates that EMAGE has an experiment suggesting the gene *bmp4* is expressed in the tissue called hindlimb. Inferences are generated through rules such as expressed(Gene, Tissue) <- emageExpressed(Gene, Tissue) which can be interpreted as saying that when EMAGE believes a gene is expressed in a tissue, then that is a reason to believe that the gene is expressed in the tissue.

Switching to using the letters G and T for genes and tissues respectively makes the first rule:

```
expressed(G, T) <- emageExpressed(G, T).
```

Thus, EMAGE can say the gene is not expressed by contradicting the first conclusion:

```
∼expressed(G, T) <- emageNotExpressed(G, T).
```

The same is true for GXD:

```
expressed(G,T) <- gxdExpressed(G, T).
∼expressed(G, T) <- gxdNotExpressed(G, T).
```

Appropriate facts may include that EMAGE believes the gene *bmp4* is expressed in a tissue called hindlimb, whereas GXD disagrees:

```
emageExpressed(bmp4, hindlimb).
gxdNotExpressed(bmp4, hindlimb).
```

If the query 'is expressed(bmp4, hindlimb) true?' is posed, the following arguments will be produced:

```
expressed(bmp4, hindlimb)
   emageExpressed(bmp4, hindlimb).
   expressed(G, T) <- emageExpressed(G, T).
```

and

```
~expressed(bmp4, hindlimb)
   gxdNotExpressed(bmp4, hindlimb).
   ~expressed(G, T) <- gxdNotExpressed(G, T).
```

Ultimately there are two arguments, with rebutting conclusions. There is currently no mechanism to resolve this conflict. At this point it is not possible to state whether or not *bmp4* is expressed in the hindlimb.

5.4.1 Resolving the conflict

Currently emageExpressed seems to magically appear. Instead imagine it is derived from information in the EMAGE database. Perhaps the information that a gene is expressed in a tissue is encoded as:

```
emageTextualAnnotation(G, T, expressed).
```

Likewise, the information that the gene is not expressed:

```
emageTextualAnnotation(G, T, absent).
```

In the case of EMAGE, there are also spatial annotations:

```
spatialAnnotation(G, T, expressed).
spatialAnnotation(G, T, absent).
```

emageExpressed may be derived via the following inferences:

```
emageExpressed(G, T) <- emageTextualAnnotation(G, T, expressed).
emageExpressed(G, T) <- spatialAnnotation(G, T, expressed).
```

Correspondingly emageNotExpressed:

```
emageNotExpressed(G, T) <- emageTextualAnnotation(G, T, absent).
emageNotExpressed(G, T) <- spatialAnnotation(G, T, absent).
```

Similar rules can be created for GXD; however, it does not have spatial annotations:

```
gxdExpressed(G, T) <- gxdTextualAnnotation(G, T, expressed).
gxdNotExpressed(G, T) <- gxdTextualAnnotation(G, T, absent).
```

Most people prefer textual annotations to spatial annotations because spatial annotations are derived using complex image processing techniques whereas textual annotations are raw expert opinion. Accordingly, most users will favour rules featuring textual annotations over rules utilising spatial annotations. To capture this, a confidence level is assigned to each rule. Rules with textual annotations will have a confidence level of *STRONG* whereas rules with spatial annotations will be assigned a lesser degree of confidence: *MODERATE*.

When inspecting the resources EMAGE suggests the gene is expressed and GXD has the opposite view:

```
spatialAnnotation(bmp4, hindlimb, expressed).
gxdTextualAnnotation(bmp4, hindlimb, absent).
```

Proposing the same query as before, one *MODERATE* argument is generated:

```
expressed(bmp4, hindlimb)
   emageExpressed(bmp4, hindlimb).
      spatialAnnotation(bmp4, hindlimb, expressed).
      emageExpressed(G, T) <- spatialAnnotation(G, T, expressed).
   expressed(G, T) <- emageExpressed(G, T).
```

and one *STRONG* argument is produced:

```
∼expressed(bmp4, hindlimb)
   gxdNotExpressed(bmp4, hindlimb).
      gxdTextualAnnotation(bmp4, hindlimb, absent).
      gxdNotExpressed(G, T) <- gxdTextualAnnotation(G, T, absent).
   ∼expressed(G, T) <- gxdNotExpressed(G, T).
```

As *STRONG* implies more confidence than *MODERATE*, the second argument wins the debate. On the basis of the currently available information and knowledge, *bmp4* is not expressed in the hindlimb.

Whilst manifestly simplified this example illustrates many of the key concepts in computational argumentation. Biological information is represented as facts, expert knowledge as inference rules. Arguments can be chained by using the conclusion of one argument as a premise in a second. The final arguments rebut one another. Arguments can have an associated strength, or preference ordering, and this can be used to resolve conflict.

However, such argumentation is of little use to biologists, for them natural language is required. In which case the final argument might read:

Evidence in GXD suggests *bmp4* is not expressed in the hindlimb: a textual annotation suggests the gene is absent from the tissue. CONFIDENCE=STRONG.

Regardless of which presentation is chosen, the real argumentation would have more arguments and associated conflict due to a richer knowledge model and an increased supply of information.

5.5 Argumentation in biology: a solution

To draw this chapter to a close, the current section will bring together some of the themes discussed previously to illuminate the solution proposed and give prominence to issues that require further consideration - this takes place in the second half of this section. The first half contains an explanation of why an existing solution from the more active argumentation use cases of medicine and law cannot be simply ‘dropped’ into place.

5.5.1 Distinguishing biology from medicine and law

Chapter 2 recounts applications of argumentation, in particular, within the fields of medicine and law. It is important to demonstrate that there are clear differences between biology and the other domains, which ensure a resolution for biology is unlike the solutions provided for the other domains.

Law, the field in which most AI-based research into argumentation has occurred, is plainly removed from biology. As Section 2.6.1 describes, the study of argumentation in law centres on the dialogues between lawyers and judges in a courtroom. A critical aspect of these dialogues is their formal and prescriptive nature.

In contrast, direct communication between biologists is neither governed nor restricted. The behaviour, thought processes, and decisions of a biologist are not limited by parliament and hundreds of years of legislation. Likewise, there is no formal mechanism for settling disputes or reaching decisions within the life sciences. Although biologists may engage in highly structured formal argumentation within journal papers, neither these papers nor the review process constitutes a debate (see the difference between monological and dialogical argumentation in Section 2.1.1). Accordingly, much of the research into AI & law - for example dialogue games modelling aspects of legal wrangling - is not relevant within the biological world.

Medicine is likewise clearly different from biology. Medicine is a safety critical domain in which mistakes carry a high penalty. Accordingly, medical researchers expend considerable effort creating and subsequently adhering to best practice - so-called *clinical guidelines*. During treatment, practising medics perform case-based reasoning to determine the correct course of action for their patients.

Again, none of these aspects is relevant to biology. Biology does not carry a significant punishment for error, thus reasoning is less robust. Whilst biologists may create best practise guidelines, *MISFISHIE*'s[5] lack of adoption highlights that they often do not follow them. Finally, biologists reason over information to determine what they can infer from that information, i.e. they reason about what to believe. In general they do not use case-based reasoning, as there are no cases, and they do not reason about actions. In contrast, case-based reasoning is the predominant form of inference within both medicine and law, and reasoning about what to do next is a common activity for medics.

Whilst the above discussion helps to explain why a solution from law, or medicine, was simply not transferred into the use case explored in this thesis, other justifications exist. Firstly, as Vreeswijk [77] and separately, Prakken and Sartor [71] suggest conflict

[5] `http://mged.sourceforge.net/misfishie/` - defines the data that should be collected when performing an *in situ* hybrdisation gene expression experiment.

resolution requires a domain specific solution, it would not be appropriate to 'borrow' a solution from another area. Instead it is necessary to understand and implement the mechanisms already used within the current field. Intuitively, this is sensible as these techniques are already likely to have been deemed appropriate by the experts in the subject. Secondly, whilst those fields have considerable research activity, and some commercial enterprise too, they do not offer publicly available products. For example, although the logic of argumentation has been applied in a number of tools within the medical domain, it is only used by the research team that developed it. At the time of writing, it is not possible to apply a toolkit or solution from either of these fields.

In contrast to the above protestations, it is perhaps worthwhile to note that the argumentation toolkit utilised during this work was implemented by a team working within medical informatics, according to the theory provided by AI & law researchers. Although the aim was to produce a general purpose toolkit, it could be claimed that the author's background inevitably effected the outcome. Accordingly, this thesis applies both a theoretical understanding, and an insight into best practice obtained from work in the medical and legal realms.

5.5.2 Outline solution

The forthcoming text switches its focus to the solution implemented in this work, and discussed in future chapters.

Figure 5.1 presents a high level overview of the work proposed in this project. It demonstrates that an argumentation system will receive as input biological information from resources (EMAGE and GXD) and the knowledge of how to interpret that information from an expert. A system could then produce biological arguments, similar to those in Section 5.4. This simplistic view is augmented in Figure 5.3.

It is an argumentation system's requirement to formalise information and knowledge that is central to Figure 5.3 - the formalism in the diagram is from the example in Section 5.4. In order to reason, the system must have its input presented in a particular manner - the exact style will depend on the toolkit used to facilitate the argumentation. In a similar vein, the system's "formal" presentation style will be not suitable for most humans. In summary, there is a need for four translations to occur:

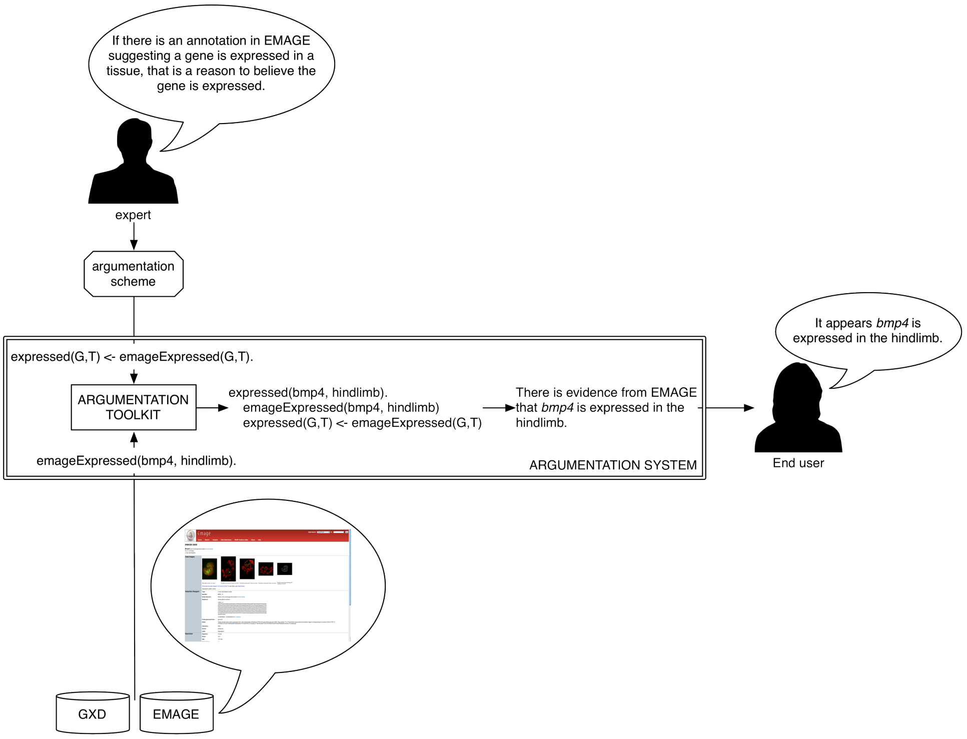

If there is an annotation in EMAGE suggesting a gene is expressed in a tissue, that is a reason to believe the gene is expressed.
expert
argumentation scheme
expressed(G,T) <- emageExpressed(G,T).
ARGUMENTATION TOOLKIT
expressed(bmp4, hindlimb).
emageExpressed(bmp4, hindlimb)
expressed(G,T) <- emageExpressed(G,T)
There is evidence from EMAGE that *bmp4* is expressed in the hindlimb.
emageExpressed(bmp4, hindlimb).
ARGUMENTATION SYSTEM
It appears *bmp4* is expressed in the hindlimb.
End user
GXD
EMAGE

Figure 5.3: An extended version of Figure 5.1 in which the need for translation between the resources, expert, argumentation system and end user is highlighted.

1. Convert the expert's natural language into a formal representation for the argumentation system;
2. Convert the programmatic representation of EMAGE's information into the argumentation system's formal representation;
3. Convert the programmatic representation of GXD's information into the argumentation system's formal representation;
4. Convert the argumentation system's formal representation of the argument into natural language for the end user.

The first translation process represents a significant obstacle and shall be the focus of the next chapter (Chapter 6); however, from Figure 5.3 it should be clear that Walton's notion of schemes (see Section 2.4.3) is used as a bridge between the expert and the computer.

The remaining three renditions will be reserved for Chapter 7, where all implementation decisions will be analysed. In addition to reporting on the translations, this chapter will feature a discussion of the toolkit used in this work, the way it was employed, and a more general review of the testbed.

An evaluation of argumentation, and the arguments generated by the system will follow in Chapter 8, with a discussion of the results, and an analysis of this activity, being reserved until Chapters 9 to 11.

5.6 Summary

The conceptual notions of arguments and argumentation were considered during this chapter. The cogitation proceeded by re-examining the use case, and continued through an examination of by which method the argumentation could work, and where the arguments originate from.

Through these explorations it was discovered that the arguments would conclude that a gene was either expressed or not expressed in a tissue, and that it must have two presentations (one for the computer and one for the human end-user). Although there are two obvious choices of which method to employ to proceed with the argumentation, both have limitations. The best option is to use expert opinion to analyse the contents of the EMAGE and GXD databases. Accordingly, the arguments must be initiated from the contents of those databases, i.e. the annotations.

There is a need to map between the communication forms of the biologists, the resources, and the argumentation toolkit. These communication gaps shall be considered in future chapters starting, in the next chapter, with a discussion of how to use the expert's biological knowledge in the argumentation process.

Chapter 6

Expert knowledge to inference rules

A brief outline of the approach taken in this thesis is given in Section 5.5. The basic idea, documented in Figure 5.3, involves capturing expert knowledge and understanding of the best ways to interpret and evaluate information in two *in situ* gene expression resources for the developmental mouse, EMAGE and GXD. The expert knowledge shall be applied to the biological information in the argument generation process. This chapter will discuss the capture of the knowledge, its documentation, and conversion into inference rules for use in the argumentation process.

Figure 6.1 concentrates on the parts of Figure 5.3 relevant to this discussion. It demonstrates that the expert wishes to communicate in natural language, but that the component of the system that uses his knowledge to argue requires its input to be written more formally in logic. An additional issue to consider, is that the expert knowledge needs to be verified, and thus read, by the expert. This forces the knowledge to be documented in natural language. If the system is to receive accurate expert knowledge that knowledge must be documented in a manner that allows fluent conversion into the system's desired format. As Figure 6.1 makes clear, ostensibly the solution is Walton's notion of schemes.

This chapter will document the process of knowledge extraction, scheme creation and translation into inference rules. In the first section (Section 6.1) the notion of argumentation schemes will be revisited, and an explanation of their suitability for this work provided. Section 6.2 follows with a description of the process used to capture

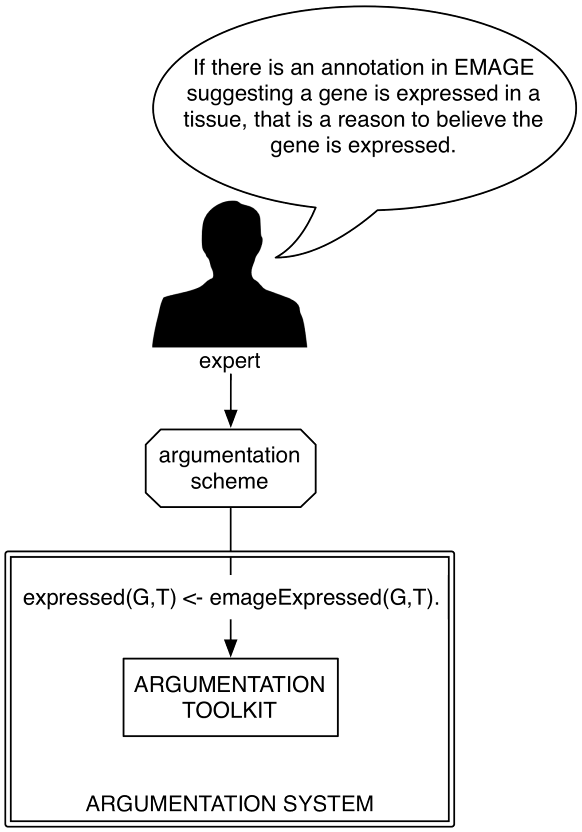

If there is an annotation in EMAGE suggesting a gene is expressed in a tissue, that is a reason to believe the gene is expressed.
expert
argumentation scheme
expressed(G,T) <- emageExpressed(G,T).
ARGUMENTATION TOOLKIT
ARGUMENTATION SYSTEM

Figure 6.1: Exploration of relationship between expert biologist and the argumentation system developed in this work - a simplified version of Figure 5.3.

the expert knowledge and convert it into schemes, before the schemes are examined in Section 6.3. The next section, Section 6.4, chronicles the process of turning the schemes into inference rules. Finally Section 6.5 summarises this chapter.

6.1 Argumentation schemes

Argumentation schemes are discussed in Section 2.4.3; here a short review will be conducted before an explanation of their suitability for this work is provided.

It is the modern interpretation of schemes, as provided by Walton [2], that is utilised here. Schemes are formed from two components. The first is the actual scheme, which is a natural language inference rule of the *modus ponens* form. Associated with the scheme is a set of questions, which aid analysis (and criticism) of the scheme's application, and arguments produced by its employment. One of the most commonly discussed schemes is Walton's argument from expert opinion, this version is from [2]:

> Source E is an expert in subject domain S containing proposition A
> E asserts that proposition A (in domain S) is true (false)
> Plausibly A may be taken to be true (false).

1. How credible is E as an expert source?
2. Is E expert in the field that A is in?
3. What did E assert that implies A?
4. Is E personally reliable as a source?
5. Is A consistent with what other experts assert?
6. Is E's assertion based on evidence?

Although there is a close link between these schemes and natural language argumentation, it is their secondary purpose of argument generation that is of interest. A scheme can be converted into a range of formal logic rules, for use in an argumentation system, by applying the method documented by Verheij [62] and discussed in Section 2.4.3.

Initially it was hoped that the above argument from expert opinion could be specialised to create a range of schemes similar to the following[1]:

> EMAGE contains an experiment suggesting that X is expressed in Y
> with confidence score Z.
> EMAGE is a leading biological resource in this field.
> Therefore we may be"Z-confident" that X is expressed in Y

Where Z represents the result of a mechanism to quantify the quality of an experiment. The exceptions and attacks on this scheme may be captured in a range of questions for example:

1. Does EMAGE truly believe that X is expressed in Y?
2. Does EMAGE really have Z confidence in this statement?
3. Is EMAGE genuinely a leading resource in the current field?

[1] Developed during a brief email exchange with Floriana Grasso.

4. What evidence does EMAGE have that X is not expressed in Y?
5. What relevant evidence does GXD contain?

Although this scheme is too simplistic, it was hoped that a range of specialised schemes, similar to this, could be developed to cover both EMAGE and GXD.

As the argumentation toolkit applied in this work (see Chapter 7 for details) resolves conflict using the strength of the arguments, calculated from the degrees of belief assigned to the component parts, it is necessary to capture these too. Ideally, there would be some form of agreed confidence measure for the biological data. However, this is not the case. It is impossible to know how accurate the data in EMAGE and GXD is. Accordingly, the degrees of belief must be associated with the rules. This requires the expert to assign a level of confidence to each scheme he creates.

6.2 Scheme creation process

The knowledge extraction took place over a series of informal meetings with the expert, Dr. Jeff Christiansen, who at the time held the role of EMAGE senior editor. These meetings had four goals:

- Understand the expert's reasoning process for analysing the data;
- Create and verify schemes documenting the expert's reasoning process;
- Assign degrees of belief to the schemes;
- Determine the best presentation of arguments.

The final goal does not relate to the work on schemes, and accordingly will be discussed in Chapters 8 and 9.

Before the first meeting took place, it was decided to minimise the technical details the expert was exposed to. To this end, he understood that there was a desire to create a range of rules that documented his reasoning processes, and a set of questions for each rule that suggested how someone could test the application of the rule. The theory of schemes was never discussed with the expert; however, he was given a rudimentary introduction to the idea of argumentation and the proposed system.

During the meetings the expert was asked to provide the link between the database facts and the inferred conclusion; the inference rule required by the argumentation toolkit. Following the meeting(s), the rest of the scheme was constructed from these rules by the analyst. Frequently the expert would discuss how to check the rules, and this knowledge was recorded as critical questions. If necessary, the expert was prompted for such knowledge. The final completed scheme, and associated questions, were shown to the expert, during a later meeting, for him to amend.

The planned outline of the meetings was as follows:

Meeting 1 High level discussion introducing gene expression concepts, and exploring the procedure carried out by the expert to review data;

Meeting 2 Concrete examination of expert's processes, leading to the creation of schemes. Beginning to consider the degrees of belief by separating schemes into groups according to importance (actual scheme creation to take place after this meeting);

Meeting 3 Verification of schemes, and concrete decisions regarding degrees of belief.

Meetings 1 and 2 went according to plan; however, in meeting 3 the expert decided that the process was not working. He proposed starting afresh, and creating new schemes by considering real examples of conflict in EMAGE. This led to the following meeting outline:

Meeting 4 Work through a range of conflicts in EMAGE with the expert suggesting how they could be resolved;

Meeting 5 Review of the schemes;

Meeting 6 Creation of degrees of belief.

Despite this "re-start" the existing schemes were not removed, unless the expert explicitly asked for that to happen. After Meeting 4, during a consultation with an evaluation expert (Dr. Gus Ferguson), it was decided that the biological expert may continue to amend the schemes indefinitely. Clearly this is impractical as it means the knowledge is impossible to fix and thus use for argumentation - for a richer discussion of the implications see Section 11.1.

To counteract the evolving schemes an artificial deadline was set with the expert given one last meeting in which he could modify the schemes. During the fifth meeting, the expert was asked to check the terminology of the schemes, but directly instructed not to add or remove any. Due to the expert's other commitments, the task of assigning degrees of belief to the schemes was restricted to one meeting - number six.

During the discussion with Dr. Ferguson, he suggested that the most appropriate way of collecting degrees of belief would centre on a card sort [152]. Hence, a series of cards were prepared each featuring a different scheme (rule only). The card sort involved the following steps:

1. Sort the cards (schemes) in order of importance[2] - several may be grouped together if they are equally important;

2. Split the cards into three groups based on confidence in them: high, medium, and low;

3. Sort the cards within their groups, again based on confidence;

4. Define a confidence range for each group, with confidence being measured as a percentage;

5. Assign each card a percentage value based on the limits for its group.

A percentage was used for the quantification of the strength, as it is a common measure that is understood by a broad group of people and is easy to normalise for use with the chosen argumentation system. The expert assigned high values 70% and above, medium 50% to 69%, and low 0% to 49%.

In the above description of the card sort, the confidence relates to the confidence a user can have in any argument produced from an application of the scheme. In effect the strength of the resulting arguments.

Chapter 11 presents a post completion review of the process used to document the expert knowledge.

[2] There is an assumption here that importance relates directly to confidence in the logic the scheme captures.

6.3 Description of schemes

This section will discuss the features, and issues, of some of the schemes created during the process described in Section 6.2 - a full listing of the schemes is provided in Appendix A. A number of insights emerged from the scheme creation work; however, they are reserved for Chapter 11.

To start the current description, a review of the terminology is provided, and that is used to fuel a discussion on the focus of the schemes. A study of the common characteristics of the schemes finishes this section.

6.3.1 Terminology and focus

The terminology, introduced in Chapter 5.3, shall be used both within the schemes and in the discussion of the schemes:

Experiment The actual research undertaken;

Result The published outcome, i.e. the researcher's interpretation of the experimental result;

Annotation The (EMAGE or GXD) editors' analysis of the experiment, including their own interpretation of the result - can be textual or spatial;

Verdict The statement that a gene is (not) expressed in a tissue, which is either stated directly in an annotation or can be derived from an annotation. The term may also be used to refer to the final conclusion drawn by the user.

6.3.2 Portrait of characteristics

The schemes can be classified into three groups: those concerned with the reliability of the underlying experiment; schemes focusing on the reliability of the annotations derived from the experiment; and schemes relating to the anatomy of the mouse.

The first group incorporates a number of subgroups that consider a range of possible issues relating to the notion of a reliable experiment. This list includes: the research team; the method used to conduct the experiment; and, the results obtained from it. The annotation group likewise has a number of subgroups focusing on spatial[3] and textual annotations[4] in addition to schemes that are applicable to both. The final group, anatomical schemes, considers how the location of the result can be used to infer new knowledge.

The schemes cover a wide range of topics and do so with differing granularity. For example the following scheme can be considered high level:

Experiment E indicates result R

Experiment $E2$ indicates result R

Experiment $E3$ indicates result $\neg R$

Two experiments indicating the same result, increases the likelihood of the result being correct

Therefore R is probably correct

1. Are E and $E2$ performed by the same lab?
2. What do other experiments show?

whereas this scheme is more focused:

The spatial annotation SA for experiment E indicates verdict V

SA is based on spatial mapping SM

SM features N number of voxels

$N < 10$

Spatial mappings with less than 10 voxels are likely to be errors in the mapping process

SA is not a good indicator that V is true

Ultimately the latter scheme was withdrawn as the expert felt that he could not state with any confidence that it was accurate. To do so would require considerable further research.

[3] The schemes in this group should apply equally to spatial annotations and textual annotations derived from spatial annotations.

[4] This category of schemes does not apply to textual annotations derived from spatial annotations.

The second scheme is a rare example of a low level scheme. However, it should be noted that it still does not feature low level biological knowledge such as knowledge of the processes behind gene expression, e.g. transcription. Instead the schemes contain practical guidance in evaluating the information in EMAGE and GXD in such a manner that little biological knowledge is required. The schemes mostly expect that their reader is familiar with the domain vernacular.

A number of schemes have no critical questions. In the case of the second scheme, it is because it was removed before the questions were considered. However, other schemes have no questions because the expert does not think they are necessary, for instance:

ISH experiment E suggests result R
The user, when examining the image from E, has confidence level C in R
The image is the actual result of the experiment, and so analysing it gives an accurate result
R is C likely to be true

The expert believes that his peers will be equally as adept at analysing an image as he is, and so he does not see any requirement to question their judgement.

During the creation process a modest number of tangential schemes were produced. For example, one scheme has the conclusion “it is very likely that the experiment used the probe”. This is never integrated into the general topic - how to analyse the contents of EMAGE/GXD. Nor is the scheme retracted as it is an accurate statement.

Numerous schemes refer to the experimental result, despite the fact that only the annotation of that result is available. For example:

ISH Experiment E produced result R

E has low pattern clarity

Experiments with low pattern clarity are very hard to analyse and so their results are not totally trustworthy

Therefore R is possibly correct

1. What does SAGE show?
2. What does Microarray show?
3. Is the equivalent result true in the previous developmental stage?
4. Is the equivalent result true in the subsequent developmental stage?

In such situations, the intuition behind the scheme can often be applied directly to the annotation. For example, as the annotation is based on the result, any conclusion implying the experiment or its result is not entirely trustworthy must mean that the annotation is not trustworthy. This is captured in:

Annotation A was based on experiment E

E cannot be trusted

If an experiment cannot be trusted, an annotation based on it cannot be trusted

Therefore A cannot be trusted

A further issue is that a number of schemes cite other resources, e.g. :

Experiment E on the mouse produces result R for gene G and tissue T

Experiment $E2$ on the zebrafish produces result R for the zebrafish equivalents of G and T

Finding the same result in multiple species provides a very good indication the result is correct

Therefore R is very likely to be correct

This scheme suggests that finding an equivalent result in another species is a good reason to trust a result for the mouse. Implementing this would entail the creation of a set of schemes to help determine when a mouse gene is equivalent to a gene from another species. A second set of schemes would be required to identify equivalent tissues across the species, with a third set vital to analyse the result of the gene expression experiments in the second species. This is an immense amount of work, and beyond the knowledge of the current expert. Accordingly, this work was not undertaken.

With regard to the argument (scheme) strengths, although the expert had a range of 0 - 100 to choose from, he grouped scores at specific points. An examination of Appendix A reveals that only 10 strengths were used: 100%, 99%, 85%, 80%, 65%, 55%, 50%, 45%, 15%, and 1%. One contributing factor to this effect was that occasionally the expert identified families of schemes that, although subtly different, were based on the same intuition. Such families were assigned the same scores.

6.4 Schemes to rules

In order for the expert knowledge to be understood by the argumentation system, it must be translated into the desired language of the argumentation toolkit employed. A description of the toolkit used in this project shall be reserved till Chapter 7. However, in order for the current discussion to take place it is necessary to state that the chosen toolkit's input form is a PROLOG-like language similar to that featured in Section 5.4.

This section will describe the process for converting schemes to inference rules, which is based on the work of Verheij discussed in Section 2.4.3.

6.4.1 Scheme to rule translation

As shown in Section 5.3 it is possible to argue either about the validity of the annotations or the validity of the verdict. In the case of the former, information about the quality of the experiment, and its subsequent inspection to form both the result and result analysis, can be used to support or attack the quality of the annotation. Ultimately, the verdict is based on the annotation, and thus arguments about it are directly transferable to the verdict. Consequently, there is little pragmatic difference between arguing about the verdict or the annotation.

Biologists use resources, such as EMAGE, to answer specific questions such as: where is *bmp4* expressed? A debate on whether one annotation is accurate is interesting. Yet, if a further five annotations exist and are known to be valid, and in support of the same verdict, the first annotation becomes almost irrelevant. To illustrate, consider the situation if the first annotation supports the other five. In this case nothing has changed as confidence in that verdict was already high. Alternatively, if the first annotation disagrees with the other five then it is unlikely to be correct, as the five are known to be valid. In conclusion, an evaluation of the validity of a single annotation must include a consideration of all other related annotations. Thus arguing about an annotation is interesting in the context of the verdict.

In this work, the conflict between annotations will be ignored until it is propagated to the verdict level.

This section will discuss the actual mechanism employed to convert the schemes into rules. The first step is to classify the rules according to an abstraction. The following categories are employed:

- reasons to trust the experiment;
- reasons to doubt the experiment;
- reasons to trust the spatial annotation;
- reasons to doubt the spatial annotation;
- reasons to trust the textual annotation;
- reasons to doubt the textual annotation;
- anatomical inferences.

Reasons to doubt/trust the experiment are themselves reasons to doubt/trust the spatial and textual annotations derived from the experiment (depicted in Figure 6.2). Following on from this, reasons to trust an annotation become reasons to trust the verdict it supports, similarly reasons to doubt the annotation are reasons to doubt the verdict. This provides the overall pattern for the rules indicating how a reason to trust an experiment can eventually become a reason to support a verdict. Yet this does not state the procedure used to convert natural language statements into the required logical statements - this is considered next.

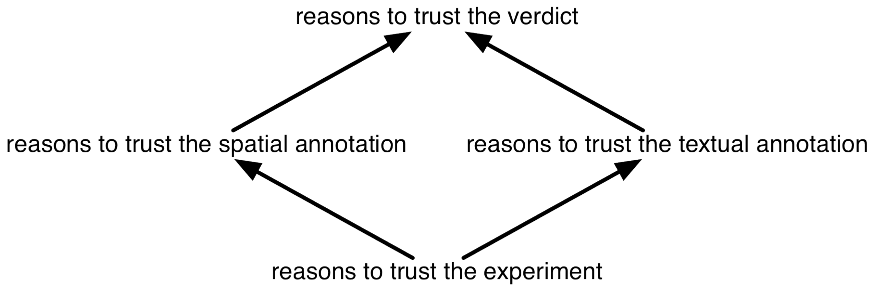

reasons to trust the verdict
reasons to trust the spatial annotation
reasons to trust the textual annotation
reasons to trust the experiment

Figure 6.2: Hierarchy of reasons to trust within an argument; there is a similar graph describing reasons to doubt.

Process

To illustrate the actual process for translating the schemes, consider the following scheme:

Spatial annotation SA suggests verdict V
Spatial annotations are a reasonable indication of expression
Therefore it is likely that V is true

1. Is SA reliable (correct)?
2. What do the textual annotations show?
3. How good is the morphological match?
4. How high is the pattern clarity?

It is a straightforward scheme that captures an intuitive idea: spatial annotations provide a reason to believe a particular verdict.

Conversion to a logical inference rule is a manual process, which requires the operative to have both a knowledge of schemes and logic. As the expert does not possess the latter, transformation was performed by the analyst.

Using the same grammar as Section 5.4: a spatial annotation (sa) indicates that a particular experiment (E) suggests a gene (G) is (not) expressed in a certain percentage

(P) of a tissue (T). L represents the precise level of expression. Accordingly, a spatial annotation can be presented as:

sa(E, G, T, L, P).

A verdict (verdict) states that a gene and tissue are associated through a level of expression:

verdict(G, T, L).

Whether or not 'spatial annotations are a reasonable indication of expression' is a decision for the user. Yet, it may be assumed, in general, users agree with this statement, otherwise there would be no reason for spatial annotations to exist. Likewise, as the aforementioned annotations are EMAGE's raison d'être, if they were not widely received it is questionable if EMAGE would exist. As such, the statement is accepted without further investigation or consideration. The reduced scheme now states that a verdict supported by a spatial annotation is likely to be true:

verdict(G, T, L) <- sa(E, G, T, L, P).

This representation is identical to one of the formats accepted by the argumentation toolkit used in Chapter 7. To include the toolkit's required expert assigned strength (65%), the following is used:

verdict(G, T, L) <- sa(E, G, T, L, P) 0.65.

Converting the natural language rule element of the scheme into a formal logic inference rule is somewhat intuitive. Transforming the critical questions is less obvious. Here the work of Verheij [62] is applied. He recounts four classes of critical questions and advises how each class may be implemented - see Section 2.4.3.

The critical questions relating to the current scheme all propose possible approaches for attacking an argument derived from the scheme. The classification of critical questions according to Verheij's categories is subjective. Therefore, these questions may be implemented by adding extra conditions to the inference rule, creating an undercutter or through the development of other rules that rebut the above inference rule. In this case, the questions all refer to existing schemes, and thus the final means is adopted.

The full list of rules can be found in Appendix B.

6.4.2 Limitations and considerations

A number of limitations and exceptions apply to the discussion about schemes and their translation into rules. These issues shall be highlighted and discussed here. Further analysis will be reserved until Chapter 11.

The first restriction is that not all of the captured schemes were implemented. Several were removed by the expert. Others were omitted because they required an appropriate new resource, and thus a new set of schemes, for which there was no allocated time.

Also discounted were schemes that required the user to analyse the experimental images. A real world system may allow the user to view these and record an opinion of them; however, only experts could do so. For the system's evaluation only two experts could be guaranteed to be available, and there was no desire to intimidate the other evaluation subjects[5] by highlighting how little they knew of the domain.

A number of practical implementation decisions had to be taken when making the rules. For example, one scheme states that when EMAGE and GXD have different textual annotations (one says the gene is expressed, the other not expressed) for the same experiment that is a reason to doubt both annotations. Although clear in meaning, this scheme does not state which method to use to decide when two experiments are the same. Nor does any other scheme. In this instance, the decision was made by comparing the citation information for experiments.

The final limitation is that it is possible to have, depending on the experimental data, a situation in which the reasons to trust an annotation are of equal strength to the reasons to doubt it. This occurs partly because of the design of the rules, and partly because of the expert assigned strengths. In this situation the outcome will depend on the semantics employed in the engine - preferred credulous or grounded. The preferred credulous semantics (see Section 2.4.6) resolves disagreement by plac-

[5] The other subjects are a mixture of Heriot-Watt University students and EMAP employees that have expertise in other areas. See Chapter 8 for more information.

ing conflicting arguments in different, equally valid, extensions. Resultantly, both arguments (trust and doubt) are justified. The grounded semantics (Section 2.4.6) builds only one extension, and is therefore sceptical in nature: neither argument will be accepted.

6.5 Summary

This chapter focuses on the biological knowledge necessary to interpret the data contained in the databases featured in the use case. It documents the process by which this knowledge was obtained, and discusses the mechanism for recording it.

That mechanism is Walton's schemes - a full list of the schemes generated in this work is provided in Appendix A. Some of the schemes, and their features are highlighted in this chapter before a discussion of the procedure for turning them into the inference rules required by the argumentation toolkit is conducted. The chapter is rounded off with a consideration of the disparities between the inference rules and the original schemes.

The quality and relevance of the schemes shall be considered during the evaluation in Chapter 8. Both the quality of the schemes, and the translation process shall be considered further in Chapter 11.

Chapter 7

Arguing over gene expression: implementation

This chapter investigates the development of two prototype systems designed to tackle the issues of inconsistent and incomplete information discussed in Chapter 4.

The opening section (Section 7.1) examines the outline requirements for a system, before the initial implementation decisions that shaped the prototypes are reviewed in Section 7.2. Section 7.2 also features an analysis of the argumentation toolkit employed in this work. Subsequently, Section 7.3 considers the architecture and design of the prototypes, with the subtle differences between the two prototypes highlighted. In Section 7.4 a brief discussion of translating information to/from the argumentation engine continues the discussion started in Chapter 6. Following on, the first and second prototypes are described in Sections 7.5 and 7.6. The chapter is completed by linking the second prototype with the evaluation carried out on it - the evaluation being the focus of Chapter 8.

7.1 Outline requirements

In order to set a realistic context for this work, and thus the requirements capture, the assumption will be that the tool created should be as close to usable in the real world as possible. To this end a pseudo situation is proposed in which the goal should be to produce a working system that could be deployed by one of the existing resources in the field in order to help their consumers resolve the issues discussed in Chapter 4.

With this in mind it is necessary to consider in what manner biologists would access such a resource. The first step is to review in what ways other biological information is accessed.

Web based publication has supplemented traditional forms, and now large resources such as GXD and EMAGE not only present a range of information to their users but additionally provide a gateway to more in depth information, such as journal articles, which are available online through different resources. Although data in the chosen resources can be accessed via programmatic means, or downloaded in a file based format, most users still approach the information via the web based interfaces. In addition to experimental information, biologists commonly utilise web based tools such as *BLAST* [153]. As such, members of the biological community are familiar with the models and requirements of web based interfaces. These points lead to the conclusion that this work should provide a web based interface.

BLAST is an intriguing tool[1]. Not because of what it does, but because of the method it employs. Unlike many resources on the web, BLAST does not attempt to provide the user with an immediate response to their query. Instead it tells the user that a response will come, but later. It provides the user with the option of waiting, or downloading the result in the future. The popularity of this tool proves that biologists are willing to accept a time delay, if there is no other choice. This is important because the fetching of data from multiple resources, the subsequent processing of that data, and finally the argumentation process collectively ensure that the response may not be instantaneous.

The goal of this work is to evaluate argumentation within a biological context, not to create the perfect argumentation tool for biology. Doing the latter would require the argumentation theory to be specialised, then implemented and optimised. Inevitably this would result in the work being mired down in the lower level details of designing a theory and framework - choosing which logic to use, deciding on what mechanism to use to implement the notion of conflict, *et cetera.* There is no evidence that such a substantial effort is worthwhile - accordingly this work shall use a pre-existing argumentation toolkit, rather than develop its own. This is in keeping with the way in which most resources would adopt argumentation. There is little likelihood of the

[1] Technically BLAST is an algorithm, which is implemented to provide a tool.

community being anything other than a consumer of a ready-made software framework. They would deploy a common toolkit that is well documented and supported by a known organisation, either commercial or open source (e.g. *Apache Software Foundation*[2]).

Furthermore, this thesis will not produce a 'complete' or 'fully-functioning' tool that will resolve all the issues discussed in Chapter 4 - doing so is beyond the resources of such a project. Yet, it is worthwhile to consider briefly what may constitute such a solution. Originally, this work started with the idea of generating an argumentation tool to assist non-expert biologists, bioinformaticians, and software agents in determining whether or not a gene is expressed. Such a product should be able to recommend a result, and provide a justification for doing so. The reasons should be comprehensible to non-expert humans and should make sense to any expert asked to verify the result. Although a range of prototypes has been developed, as the subsequent chapters make clear, this line of work requires further research. Argudas, Chapter 12, presents a different view of a 'complete system', because it caters exclusively for experts. Essentially, the foundation of these two systems is the same; however, the argumentation performed and the visualisation of that argumentation is different. Further unification may be possible as the technology develops.

Detailed requirements for the actual tool were produced in collaboration with the expert biologist in a process that was aligned to the scheme creation task. Requirements were discussed with him, and evolved through the development of the first prototype (Section 7.5), on which the expert provided comment.

7.2 Test bed

This section examines the actual resources and tools used within these prototypes; more details are presented in Section 7.3 where the architecture is discussed.

The prototypes are built upon the data provided by two online resources, EMAGE and GXD. A description of these resources is provided in Chapter 3. Section 7.2.1 discusses why these two resources were selected in preference to the various alternative services that are available.

[2] www.apache.org

Furthermore, Section 7.2.1 relates the reasons for selecting the biological expert. He was a crucial part of the prototype implementation, as he provided both the expert knowledge necessary to create arguments, and guidance for the development of the system, in particular the user interface.

The final main component of the system is the argumentation toolkit. The following section, Section 7.2.1, declares and justifies the selection of a particular toolkit before Section 7.2.2 describes and evaluates the chosen toolkit.

The argumentation toolkit is implemented in JAVA, thus it made sense to continue with its use for the development of the web application. JAVA is augmented with HTML and, in the second prototype, JAVASCRIPT. The web applications are served by an *Apache Tomcat*[3] installation. Details of the two prototypes developed during this work will follow in Sections 7.5 and 7.6.

7.2.1 Initial decisions

Decisions made before the implementation began had major implications for the work undertaken. The main decisions will be recounted and championed here.

EMAGE and GXD

As illustrated in Section 3.5, there is a wide range of resources publishing the results of gene expression data for the developmental mouse. Furthermore, Chapter 4 advises that it is necessary to use many resources to reduce the level of incomplete information. Despite this, only two resources are featured: EMAGE and GXD. Furthermore, irrespective of these resources containing many different types of gene expression experiment, only one (*in situ* hybridisation) is used.

There is no doubt that this is a limitation that restricts the range and number of arguments that can be produced. Crucially though, this limitation does not affect the validity of the arguments that are generated. However, prompting from the expert, see Section 7.5.2 for details, ensured that a third resource was added to the second prototype.

The overarching goal of this work is to investigate argumentation, not produce

[3] `http://tomcat.apache.org/`

a comprehensive, fully supported, system. The prototypes being no more than a mechanism through which argumentation in biology can be explored. As such, the prototypes need to have enough functionality to allow a user's reaction to arguments and argumentation to be tested. There is no requirement for a complete system, despite the obvious desirability of one.

In addition, the full inclusion of other resources requires fresh expert knowledge and thus new experts. *In situ* hybridisation for the developmental mouse was selected as the domain for the use case not only because it illustrates the problem well, but additionally because locally based experts are readily available. The same cannot be said for other experimental types.

As the expert worked for EMAGE, the EMAGE database was used as a data source. EMAGE has a close relationship with GXD, and thus it is used as the second resource. Moreover, the author already had a number of years of experience using these resources as a client. There is no reason why competing resources could not have been used instead; however, in this case, the chosen resources are more amenable.

Selecting EMAGE opens up the idea of spatial annotations, as many resources do not provide these. However, this results in another decision point as discussed next.

Textual annotations derived from spatial annotations

EMAGE's spatial annotations are the result of someone (most likely an EMAGE editor) mapping the experimental result (image) onto the appropriate EMAGE 3D model. Such annotations are commonly displayed to the user as a 2D image with particular colours indicating the different levels of expression.

Although the three EMAGE editors could interpret these images, the remaining likely evaluation participants would not be able to do so due to a lack of expertise. As these images are a human interpretation of the experimental result (image) an expert would prefer to see, and analyse for himself/herself, the real experimental result. Accordingly, there seems to be no tangible benefit to using the spatial annotation images, hence they are never employed.

The obvious alternative is to try to use the spatial annotations automatically. No software is publicly available to enable interpretation of the spatial annotation images. However, EMAGE developed an experimental technique to automatically analyse the spatial annotations and turn them back into textual annotations (so-called *Textual annotations derived from spatial annotations*). This process was used for a limited time because it required an enormous amount of effort to upgrade the 3D models to make it work correctly. The outcome was that only a small number of these annotations exist.

For the sake of clarity textual annotations derived from spatial annotations will be labelled as spatial annotations, and the schemes for spatial annotations will be applied. The latter decision, to apply the spatial annotation schemes, is in keeping with the views of the expert.

Argumentation engine

As Chapter 2 demonstrates, many different types of argumentation software exist. However, this work requires a tool that generates and evaluates arguments. The prototype is intended to be a web-based application. For this reason the argumentation toolkit must be a robust tool that can be integrated readily into applications.

At the onset of this work few tools that met these requirements were available. Many different theories for argumentation have been proposed, but few have a resilient implementation that can be integrated freely into another application. For example, Gordon [101] produced an implementation of his theory for The Pleadings Game, but it is not available publicly. *Oscar* is an implementation from Pollock [53] which is available for download. It is a complex set of LISP code designed to test the academic theories of its creator, and thus is not designed for third party use. OSCAR is not supported, thus the technical issues encountered when attempting to use it could not be resolved. The original argumentation engine concept and theory was produced by Dung [31], but it did not include an implementation.

The lack of a generally usable argumentation toolkit prompted the creation of the EU funded *Argumentation Services Platform with Integrated Components* (ASPIC) project. ASPIC summarise the state of available argumentation toolkits before their project as follows [154]:

> Existing implementations seem more or less simple proof of concept systems. None are robust enough to be standard, reusable and portable for deployment in applications. (page 5)

The outcome of the ASPIC project was the one tool that was readily available, easily integrated into a web development, free to use, and (at that time) actively developed. Hence it is the tool selected for use in this project. A thorough discussion of this toolkit can be found subsequently in Section 7.2.2.

7.2.2 ASPIC argumentation toolkit

Argumentation Services Platform with Integrated Components (ASPIC) was an EU FP6 project which brought together theorists and those interested in the practical application of argumentation in medicine and law. Fox et. al. [155] describe the goals of the project as follows:

> ASPIC is an EU funded research project (www.argumentation.org) whose goals are to develop a theoretical consensus for four roles of argument (in inference, decision-making, dialogue and learning) and validate it within the context of a general software agent.

Disappointingly, the project's web presence (`www.argumentation.org`) is no longer active[4]. It contained a simpler description of the project using a two item list. The second item in the list[5] read as follows:

> To develop efficient proof procedures and software component implementations of these models for deployment in real-world applications.

From these similar quotes it should be clear that the project focused on the creation, and implementation, of a theory for argumentation in four aspects of AI: argument inference; dialogue; decision making; and machine learning.

The inference component, called the *argumentation engine*, is an instantiation of Dung's theory. Originally it is described by Caminada and Amgoud [156]; though, full details are reserved for an ASPIC project deliverable [157]. This component creates arguments, and determines if they are justified according to Dung's semantics. The engine is written in JAVA, and designed to be used by the other three tools.

[4] A far smaller site exists at `http://aspic.acl.icnet.uk`

[5] Formerly available at `www.argumentation.org/goals.html`. A screen capture of this page can be seen in Appendix C.

The decision making component, takes a list of justified arguments provided by the engine, and then applies decision making criteria. The result is a preferred decision option. The dialogue tool deals with conflict resolution dialogues (one subclass of persuasion dialogue), and again uses the argumentation engine to find appropriate arguments that can be used by a player in the dialogue. Arguments are used to restrict the search-space when performing machine learning. As the engine is relied upon by the other three areas of activity, it is the most developed.

The effectiveness of these tools is illustrated by the project's demonstration applications, which include deciding on the suitability of organs for transplant [120].

ASPIC argumentation engine

The ASPIC argumentation engine is designed to create and evaluate arguments. It does this using facts and rules supplied by the user. Argument creation is a process of abduction, in which backward chaining is used to build an argument starting from the conclusion and working backwards, through facts and rules, until every premise of every rule is instantiated with a fact. Considering the facts and rules presented in Figure 7.1, the argument **P1** (from Figure 7.3) is built by following the process outlined in Figure 7.2[6].

The strength of an individual argument is assessed using either the weakest link or last link approach (Section 2.4.5). The strength, to which the weakest/last link algorithm is applied, comes from the user's assignment of a number to the facts and/or rules. This number, called the *degree of belief*, is a real number between 0 and 1 - with 0 implying definitely false, and 1 definitely true. In Figure 7.1 only the rules **r1** to **r5** have an assignment. The other rules, and all facts, take the default value of 1.

Arguments can be evaluated as *defeated* (false), *undefeated* (true) or *unknown* (no status assigned) using either Dung's grounded semantics (Section 2.4.6) or a credulous version of Dung's preferred semantics (Section 2.4.6).

[6] Figure 7.2 is a cleaned up version of log generated by the ASPIC argumentation engine.

```
facts:
        a. b. c. d. e.
rules:
        [r1] x <- a 0.4.
        [r2] x <- b 0.5.
        [r3] x <- c 0.7.
        [r4] ~ x <- d 0.6.
        [r5] w <- d 0.7.
        [r6] y <- x.
        [r7] ~ y <- w.
```

Figure 7.1: Example rules and facts that may be used with the ASPIC argumentation engine.

```
searching for arguments for: y
  found rule: [r6] y <- x.
    searching for arguments for: x
      found rule: [r1] x <- a 0.4.
        searching for arguments for: a
          found fact: a
          Arg1 = a
  Arg3 = x 0.4. (from Arg1 and r1)
Arg5 = y 0.4. (from Arg3 and r6)
```

Figure 7.2: Description of the process undertaken to build an argument for y using the facts and rules from Figure 7.1.

The process of generating and evaluating arguments is wrapped within a dialogue game. Two virtual players respond to a query from the user - the query should match a potential conclusion of an argument[7]. The first player (ASPIC calls it ‘PRO’) attempts to generate arguments whose conclusion matches the query, and thus supports it, helping to prove the query can be evaluated as true. The second player (called

[7] If there is no possibility of an argument with an identical conclusion to the query being constructed argumentation is not possible.

'OPP') tries to stop the first - by producing arguments that rebut either the query or an argument produced by player one. If player two succeeds with an argument, the first player can attempt to rebut that argument or develop a new argument to support the query.

At each stage in the dialogue game, an argument is generated and then compared to the existing arguments. If the new argument is stronger (has a higher degree of belief) than the opposing argument(s), it temporarily defeats the opposing argument(s). The defeated argument(s) may be reinstated if the attacker is later opposed by an even stronger argument.

Eventually one of the players will be unable to defeat the arguments of the other player. This signals the end of the game, at which point the arguments, and any status assigned to them, may be examined by the user. The programmatic interface to the engine provides a range of display mechanisms that allow the arguments to be reproduced both textually and diagrammatically.

Due to the design of the system, the strongest argument always decides which side wins the debate. If the first player has the strongest argument, the argument with the highest degree of belief, then it will reinstate many (perhaps most) of the first player's arguments. Consequently, the majority of the first player's arguments will be true, and the second player's false. The situation is reversed if the second player has the strongest argument. If both players use an equally strong argument, then the choice of semantics decides what happens. Grounded semantics are sceptical, thus all arguments are defeated. Alternatively, preferred credulous semantics accepts the arguments as equally valid.

To illustrate[8] the above discussion, assume a user provides the ASPIC argumentation toolkit with the knowledge from Figure 7.1, and then asks ASPIC for arguments supporting the conclusion y. Initially, PRO develops **P1** (see Figure 7.3) as outlined in Figure 7.2. Immediately, OPP looks for arguments for ∼y and generates **O2**. **O2** is stronger than **P1**, and thus defeats it. Now PRO attempts to rebut **O2** by rebutting ∼y with **P2**, but **P2** is too weak. PRO then develops **P3**, which has the same strength as **O2**. What happens next depends on the semantics chosen for the game. If preferred credulous PRO succeeds; however, if grounded (sceptical) PRO fails.

[8] The following examples are simplified in order to improve clarity.

Assuming PRO succeeds, OPP then looks for ways to defeat **P3**. OPP tries to generate new arguments to rebut y, but cannot. OPP then attempts to rebut the sub-argument for x, in doing so OPP generates the argument **O1**. Both this argument, and **P3**'s sub-argument for x (i.e. **P4**) have the same strength. With credulous semantics both arguments are successful, and thus the attack fails. OPP has no other option for rebutting x, and so attempts to rebut the sub-argument for c. However, this argument is based on the fact c, which has a degree of belief of 1, and thus cannot be defeated. OPP is unable to defeat **P3**, thus **P1** is reinstated. Having failed to rebut the conclusion of **P1**, OPP will try to attack the sub-arguments of **P1**; these attacks follow the same lines as the attacks on **P3** and therefore fail in the same ways. PRO now restarts the process by putting forward the argument **P2**. Following that argumentation process, PRO starts another debate with **P3**. As preferred *credulous* semantics are employed **P1**, **P2**, **P3** and **O2** are all undefeated.

With grounded (sceptical) semantics, PRO cannot defeat **O2** with **P3**. PRO has no more arguments with which to rebut the conclusion (∼y), and so attempts to rebut the sub-argument for w. PRO cannot build an argument for ∼w, and thus attempts to rebut **O2**'s sub-argument for e. PRO is unable to develop an argument for ∼e. There is nothing else for PRO to attack, accordingly OPP wins. PRO starts a new line of argumentation by using **P2** as the initial argument. In the same way OPP wins this too. Finally, PRO loses a debate by starting with **P3**. As grounded semantics are sceptical in nature, and PRO's best argument has the same strength as OPP's best argument, all the arguments for y and ∼y in Figure 7.3 are defeated.

For more information on the argumentation engine see Appendix D.

PRO arguments:

P1 - strength is 0.4 because of [r1].

```
y
   [r6] y <- x.
      [r1] x <- a.
         a.
```

P2 - strength is 0.5 because of [r2].

```
y
   [r6] y <- x.
      [r2] x <- b.
         b.
```

P3 - strength is 0.7 because of [r3].

```
y
   [r6] y <- x.
      [r3] x <- c.
         c.
```

P4 - strength is 0.7 because of [r3].

```
x
   [r3] x <- c.
      c.
```

OPP arguments:

O1 - strength is 0.6 because of [r4].

```
~x
   [r4] ~x <- d.
      d.
```

O2 - strength is 0.7 because of [r5].

```
~y
   [r7] ~y <- w.
      [r5] w <- e.
         w.
```

Figure 7.3: Some arguments generated from the knowledge in Figure 7.1.

Evaluation of ASPIC toolkit

In many ways the ASPIC argumentation toolkit is exactly the type of product desired. It is designed to be plugged into software architectures, and is written in JAVA. ASPIC provides the full framework, so there is no need to become lost in lower level questions such as "how should arguments be represented?" - the toolkit comes with an implementation.

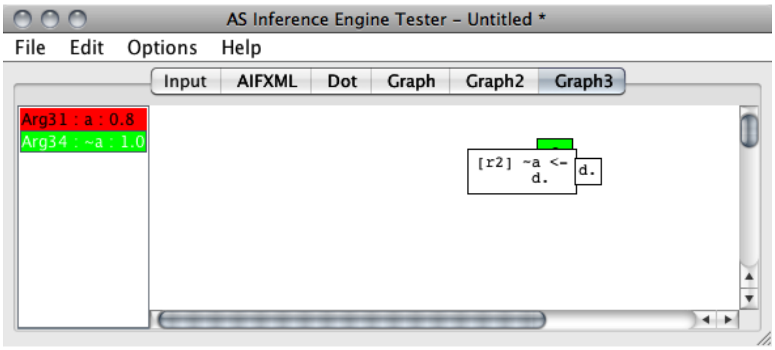
In other aspects the use of the ASPIC component is less than ideal. It provides two styles of argument presentation: visual and textual. Visual presentation of arguments is based on the standard method of displaying an argument as a directed graph. ASPIC does this by producing output in a format that can be read by other toolkits, such as *DOT*[9]. ASPIC's implementers provide a client to their tool that demonstrates these presentations - none of which is suitable. This is partially because the argument layouts are often badly displayed with arguments obscuring one another[10] (e.g. see Figure 7.4), and partly because some display graphs are too esoteric (e.g. Figure 7.5).

Textual presentation is little better. A typical example can be seen in Figure 7.6. In this figure there are three arguments, each starting with the line 'UNDEFEATED'. There are three types of output in a typical argument.

The first is a natural language statement, such as "The experiment has an assay score of 2". Such statements are defined by the person programming with the argumentation engine.

[9] `www.graphviz.org`

[10] It is possible to manually rearrange the boxes to make the content visible.



```
Arg31 : a : 0.8
Arg34 : ~a : 1.0
```

```
[r2] ~a <-
      d.
d.
```

Figure 7.4: Screenshot from the ASPIC argumentation engine test interface - some of the argument graphs do not display clearly.

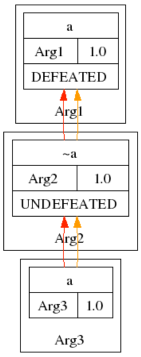
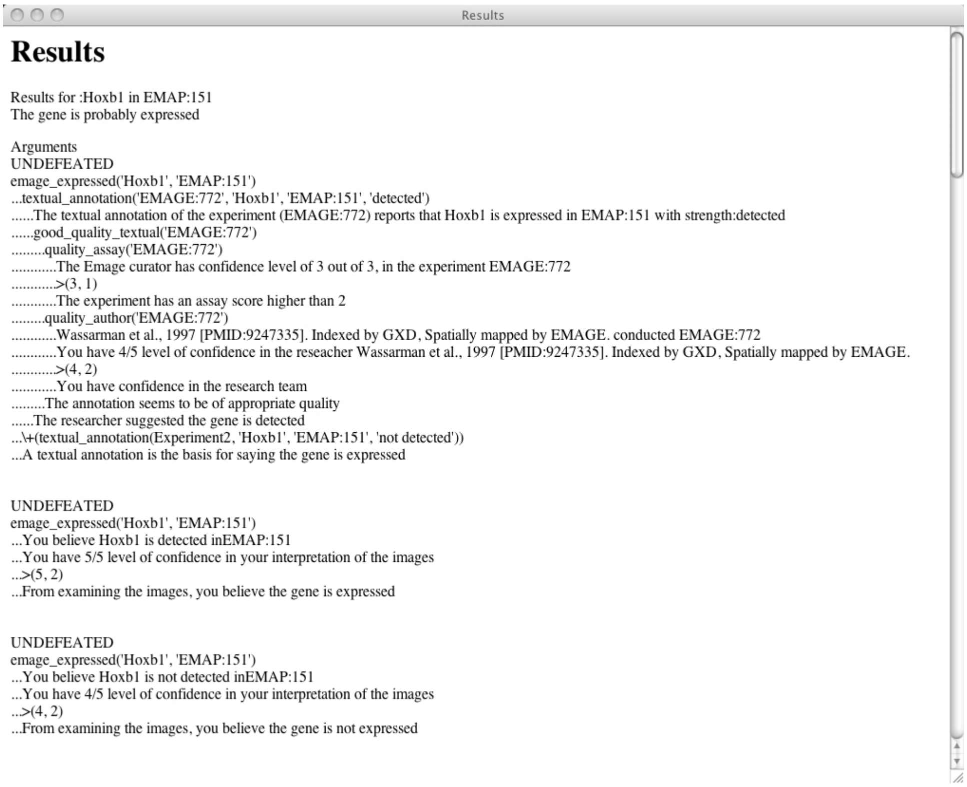


Figure 7.5: Example of a DOT graph produced by the ASPIC argumentation engine.

An example of the second type is “emage_expressed(‘Hoxb1’, ‘EMAP:151’)”. This is the *claim* of the argument, written in ASPIC’s logical representation. Initially, this claim had to be displayed in the logical form; however, eventually the ASPIC developers provided a mechanism to associate a natural language statement with a claim, which could be presented instead.

The final type of output in Figure 7.6 is similar to “>(3,1)”. This reports the result of a built-in operation, in this case a numerical comparison in which the operation asks if the first number is greater than the second. The output is controlled by the engine’s developers.

The mix of styles is confusing and difficult to read. Regrettably, all the initial default presentation styles suffered from the same limitation. By the time the ASPIC developers provided a suitable alternative, discussions with the expert biologist had yielded a mechanism for use in this work.



```
Results

Results for :Hoxb1 in EMAP:151
The gene is probably expressed

Arguments
UNDEFEATED
emage_expressed('Hoxb1', 'EMAP:151')
...textual_annotation('EMAGE:772', 'Hoxb1', 'EMAP:151', 'detected')
......The textual annotation of the experiment (EMAGE:772) reports that Hoxb1 is expressed in EMAP:151 with strength:detected
......good_quality_textual('EMAGE:772')
.........quality_assay('EMAGE:772')
............The Emage curator has confidence level of 3 out of 3, in the experiment EMAGE:772
............>(3, 1)
............The experiment has an assay score higher than 2
.........quality_author('EMAGE:772')
............Wassarman et al., 1997 [PMID:9247335]. Indexed by GXD, Spatially mapped by EMAGE. conducted EMAGE:772
............You have 4/5 level of confidence in the reseacher Wassarman et al., 1997 [PMID:9247335]. Indexed by GXD, Spatially mapped by EMAGE.
............>(4, 2)
............You have confidence in the research team
.........The annotation seems to be of appropriate quality
......The researcher suggested the gene is detected
...\+(textual_annotation(Experiment2, 'Hoxb1', 'EMAP:151', 'not detected'))
...A textual annotation is the basis for saying the gene is expressed


UNDEFEATED
emage_expressed('Hoxb1', 'EMAP:151')
...You believe Hoxb1 is detected inEMAP:151
...You have 5/5 level of confidence in your interpretation of the images
...>(5, 2)
...From examining the images, you believe the gene is expressed


UNDEFEATED
emage_expressed('Hoxb1', 'EMAP:151')
...You believe Hoxb1 is not detected inEMAP:151
...You have 4/5 level of confidence in your interpretation of the images
...>(4, 2)
...From examining the images, you believe the gene is not expressed
```

Figure 7.6: Results page of the first prototype.

The fact that the ASPIC argumentation engine provides all five levels of Prakken's framework was considered initially to be entirely positive. However, there are unfortunate side-effects. It means that a user is forced to use ASPIC's chosen logical language, ASPIC's notion of argument strength, and ASPIC's choice of mechanism for comparing arguments[11]. The use of a first order formal logic means that the notion of schemes is required to bridge the communication gap between the engine and the biological expert. The required argument strength, and lack of a suitable mechanism to assign a degree of belief to biological facts, necessitates that the expert rules have to be assigned a confidence value. As the strengths ultimately need to be a number between 0 and 1 it is a rational option to try and capture it in a similar way. Yet as reported in Chapters 6 and 8 this was not entirely successful. The availability of only weakest link and last link approaches for calculating argument strength seems acceptable as the former is believed to be intuitive and easy to explain to biologists,

[11] The ASPIC toolkit is provided in a modular form so that a user can program alternatives; however, the code is not open source which makes this difficult.

making it a natural choice.

The final concern with ASPIC's software was that it was still in development with the first version of the argumentation engine used being 0.4.3, and the final version 0.4.10. As it never reached production quality, there are some minor errors and questions over the ASPIC toolkit's efficiency (see Appendix D). Despite the obvious faults the ASPIC argumentation engine was the most suitable candidate.

7.3 Architecture and design

Two prototype systems were developed, both designed around the same architecture. This section shall explore that architecture and discuss some of its limitations.

Figure 7.7 presents the architecture of both prototype systems. It is simplified in the sense that it hides the lower level details and organisation; however, such information is not currently obligatory.

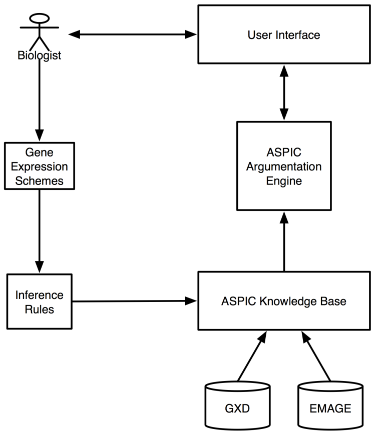


Figure 7.7: Basic architecture for prototypes.

The system starts and ends with a biologist - not necessarily the same one. The stick figure, in Figure 7.7, represents both the biological expert and the end user. Considering the former, his knowledge is captured and converted into inference rules, via the temporary step of argument schemes. The inference rules are loaded into the ASPIC toolkit's knowledge base.

When an end user specifies a gene-tissue pair in which (s)he is interested, via the user interface, the system pulls the relevant[12] information from EMAGE and GXD, processes it, and stores it in the knowledge base. A query is sent to the ASPIC argumentation engine, and the resulting arguments are presented to the end user via the user interface.

7.3.1 A more detailed examination

From the above description of Figure 7.7 it is clear that the diagram, in order to aid clarity, is missing several components. Considering the right hand side of the diagram only - where the user selects a gene-tissue pair of interest - a more thorough exploration of the components and how they interact is provided in Figure 7.8.

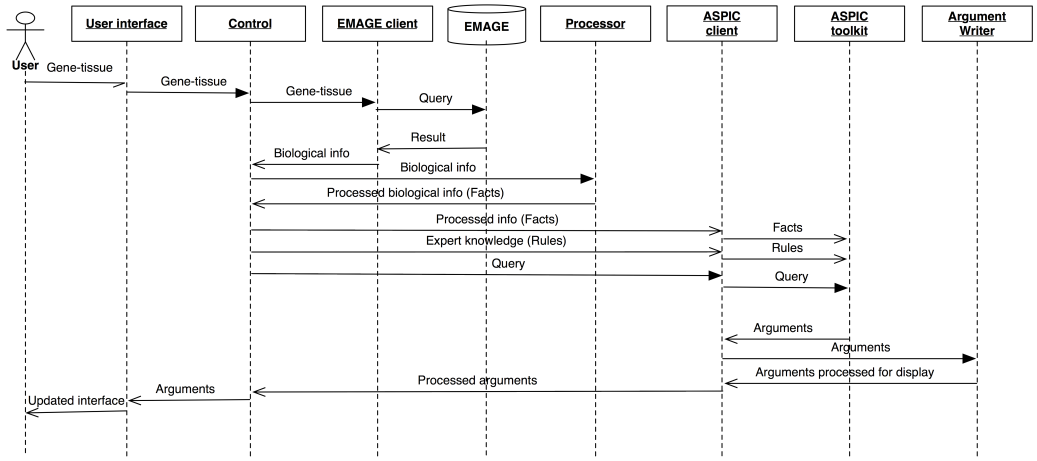


Figure 7.8: Sequence diagram of underlying activities in the prototypes. The various ASPIC components are grouped under the title 'ASPIC toolkit' and the components related to the GXD database and local cache are missing. Three different mechanisms for displaying arguments exist; however, these are all represented by the 'Argument Writer'.

As the project aims to create a web application the standard Model View Controller model of design is preferred. In Figure 7.8 the *View* is represented by the *User interface*, likewise *Controller* becomes *Control*. The remaining elements are part of the *Model*, or external resources such as EMAGE.

The flow of information through the system is controlled by the *Control* object. It starts by taking the user's desired gene-tissue selection and arranges for biological

[12] Defined as all *in situ* experimental information available for the gene-tissue combination.

information to be collected from the underlying resources. For brevity's sake Figure 7.8 only considers EMAGE, but all other data sources will be accessed at this time, e.g. GXD. *Control* interacts with a client specific to each resource. It is the job of the client to extract information from the resource it represents, and then convert that information into the data model for the project.

Information on which authors conducted the relevant experiments, and in which journals they were published, is presented to the user. The expectation is that the user will have a level of trust they associate with each journal and researcher, and thus the biological information can be treated accordingly. This idea comes from two schemes provided by the expert, and is viewed as a means by which a degree of belief may be associated with a small number of facts.

There are a number of subtle differences in the workflows between the two prototypes (summarised in Table 7.1). The workflow for prototype one was based on the schemes and understanding held at the time. Subsequently, the expert abandoned one scheme (it asked the user to evaluate the quality of the experimental results). Additionally, the expert added a scheme relating to the user's confidence in the resources (EMAGE and GXD). These changes, in conjunction with the feedback the expert provided for the first prototype (see Section 7.5.2), resulted in prototype two having an amended workflow and appearance.

Prototype one asks for the information - about the journals and researchers - during the second step; in the first step only the gene-tissue pair is requested. Prototype two is different: it asks for confidence in the journals and the gene-tissue in the first step, leaving only questions regarding confidence in the researcher for the second step.

Another difference between the prototypes is that they ask disparate additional questions. In particular, the second prototype asks for confidence in the resources, when the first does not. Moreover, the first prototype asks the user to evaluate the relevant experimental results. This was deemed unnecessary for the second prototype. The majority of evaluation subjects (see Chapter 8) could not perform this task; hence there was no worthwhile reason to retain this functionality in the second prototype. Furthermore, by not adding this question, the prototype becomes less daunting for non-expert users who made up the bulk of the evaluation subjects.

Once the appropriate biological and trust information has been received by the

Control object, it asks the *Processor* to convert the information into the format desired by the ASPIC argumentation engine - creating the so-called *facts*. The facts are sent to the ASPIC knowledge base (represented in Figure 7.8 by *ASPIC toolkit*) via the *ASPIC client*. This client is a collection of helper methods, specific to this work, to make using the toolkit easier.

Control loads the expert knowledge into the knowledge base, and then sends the query to *ASPIC toolkit* via its client. The ASPIC argumentation engine, tries to prove the query by conducting an argumentation game. Arguments generated during the game are sent to the *ASPIC client*, which converts them into a human readable form by using the *Argument Writer* before sending the arguments back through *Control* to the user.

Three different display forms were experimented with during this work - all three are represented by the abstract notion of the Argument Writer.

The components represented by *ASPIC toolkit* were the work of the ASPIC project. Similarly, the *EMAGE* database was the work of EMAP. The EMAP and ASPIC clients, in addition to all other components in Figure 7.8, are part of the prototype exercise.

Step	**Prototype 1**	**Prototype 2**
1	Asks for gene and tissue	Asks for gene and tissue, and confidence in journals and resources
2	Asks for confidence in researchers, journals, and the experimental results	Asks for confidence in researchers
3	Displays output	Displays output

Table 7.1: Summary of prototype one and two workflows.

7.3.2 Limitations and considerations

The above architecture presents something of an idealised view. Whilst creating the real system the following challenges were encountered.

Foremost, the process of accessing the information is not without problems. In particular, contacting the resources and pulling the information takes time. Additionally, the time difference between the east coast of the USA (home of GXD) and the UK means that GXD's direct SQL connections may be closed during the UK's working day to allow background maintenance processes to run. Such concerns could be resolved by the use of a local cache; however, maintaining such a repository would come at a considerable cost as both resources are constantly updating their content.

Despite the reluctance to use a cache, the list of spatial annotations is stored in one because that information is not available through the EMAGE web service. As the process of creating these annotations is suspended, this list will not change, and no update is required. Direct access to the EMAGE database would resolve this inconsistency. However, direct access would not provide, as readily, all the information required by the schemes. In particular, one group of schemes uses the notion of propagation to infer new annotations (e.g. if a gene is expressed in a paw, because that is *part-of* the limb, the gene is also expressed in the limb). This requires knowledge of the parent-child relationships within each stage of the developmental mouse, which are stored in EMAGE's database, and the grandparent-child, great grandparent-child, *et cetera* relationships that are not contained explicitly in the database. Obtaining this information for the entire length of the path (from root note to leaf node) would require multiple queries via the web service, or a database query using transitive closure. Yet, in this case, with a local database already in existence, it is logical to calculate all the relationships once and store them locally so they can be applied for both EMAGE and GXD.

7.4 Translating between argumentation engine and external entities

Chapter 6 discusses the conversions necessary for the argumentation toolkit to communicate with the expert biologist, through the staging post of schemes. In particular, Section 6.4 reviews the process for translating between schemes and the argumentation toolkit.

Section 6.4 demonstrates how the scheme:

Spatial annotation SA suggests verdict V
Spatial annotations are a reasonable indication of expression
Therefore it is likely that V is true

can be converted into the following PROLOG-like logic:

```
verdict(G, T, L) <- sa(E, G, T, L, P) 0.65.
```

However, for the rule to be used by the ASPIC toolkit, the rule must be entered into the knowledge base. The easiest way of doing that is through the JAVA programmatic interface. Here the following is used to describe the rule:

```
new Rule(new Term(“verdict”, new Variable(“E”), new Variable(“G”),
new Variable(“T”), new Variable(“L”), new Variable(“P”)), new ElementList(new
Term(“sa”, new Variable(“E”), new Variable(“G”), new Variable(“T”),
new Variable(“L”), new Variable(“P”))), 0.65);
```

7.4.1 Further conversions

Referring back to Figure 5.3, from Chapter 5, it is noticeable that two other translations exist.

The first is the translation between the resources and the ASPIC argumentation toolkit. This happens as a two stage process - as outlined in Figure 7.8. In stage one the disparate data models of the resources are unified. The second step involves converting the application's data model to the one required by the argumentation toolkit. This stage of the process is similar to the above process for converting the expert's knowledge. A two stage model is used to increase the flexibility of the system.

The conversion of arguments from their logical representation in the engine into something a user can read is the second translation. Arguments are formed from a conclusion, and multiple rules/premises; because arguments can be chained, there may be many levels of arguments to display. Furthermore there are three types of information stored for each premise/rule[13] in an argument: a logical representation, a very short textual representation, and a longer textual description. In this work,

[13] Built-in predicates only have one form: a logical one.

only the textual description was desired - it was based on the schemes produced by the expert. Accordingly, displaying an argument involves using the argumentation engine's API to navigate through each argument, presenting only the desired elements. As this is largely a programming exercise, details will not be provided.

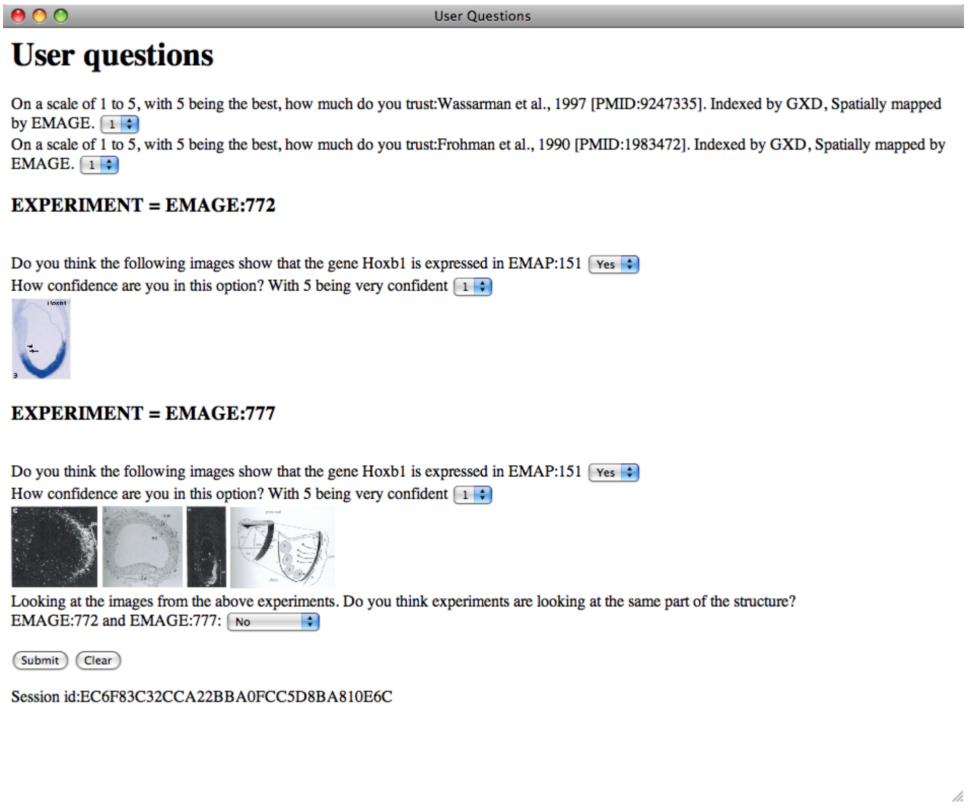
7.5 Prototype one

During the process of scheme generation (see Chapter 6) the first prototype was developed and shown to the expert in meeting 5. The system served dual purposes. Firstly to help the expert visualise the system and the relationship between it and his schemes, and secondly to gather the expert's opinion of what the system should look like. In particular, the best mechanism for the presentation of arguments.

7.5.1 Walk through

The first prototype is a basic application, with a user interface utilising HTML forms. The first screen (not shown) asks the user for their choice of gene and tissue.

Once the biological data has been retrieved from the resources, the second screen asks the user for their opinion of that data (e.g. Figure 7.9). Information sought includes the user's analysis of the results, and the user's confidence in the researchers.



User questions

On a scale of 1 to 5, with 5 being the best, how much do you trust:Wassarman et al., 1997 [PMID:9247335]. Indexed by GXD, Spatially mapped by EMAGE. 1

On a scale of 1 to 5, with 5 being the best, how much do you trust:Frohman et al., 1990 [PMID:1983472]. Indexed by GXD, Spatially mapped by EMAGE. 1

EXPERIMENT = EMAGE:772

Do you think the following images show that the gene Hoxb1 is expressed in EMAP:151 Yes
How confidence are you in this option? With 5 being very confident 1

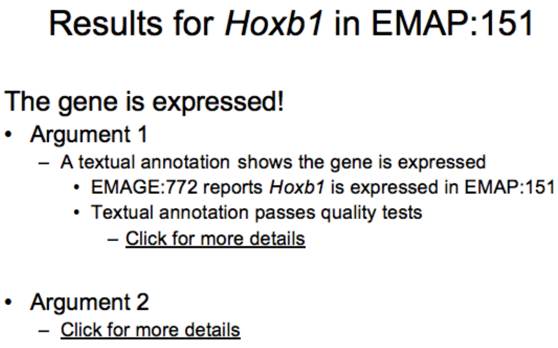
EXPERIMENT = EMAGE:777

Do you think the following images show that the gene Hoxb1 is expressed in EMAP:151 Yes
How confidence are you in this option? With 5 being very confident 1

Looking at the images from the above experiments. Do you think experiments are looking at the same part of the structure?
EMAGE:772 and EMAGE:777: No

Session id:EC6F83C32CCA22BBA0FCC5D8BA810E6C

Figure 7.9: Prototype one: asking if the user can rate the images and award a degree of confidence in the researchers who performed the relevant experiments.



Results for *Hoxb1* in EMAP:151

The gene is expressed!

- Argument 1
  - A textual annotation shows the gene is expressed
    - EMAGE:772 reports *Hoxb1* is expressed in EMAP:151
    - Textual annotation passes quality tests
      - Click for more details
- Argument 2
  - Click for more details

Figure 7.10: Possible alternative argument display mechanism for prototype one.

The third and final page (see Figure 7.6) presents the arguments generated from the biological information.

7.5.2 Initial feedback

During the discussion with the expert biologist, his opinion of the system was sought, including his views on the presentation of arguments. As discussed in Section 7.2.2, the default ASPIC presentations are not suitable for the intended user group. Subsequently a range of alternatives was presented to the expert, e.g. Figure 7.10 uses natural language and the notion of bullet points to hide more detailed information until the biologist wishes to drill down into it. Other styles shown to the expert included one based on the standard tree-based visualisation, such as Figure 7.11.

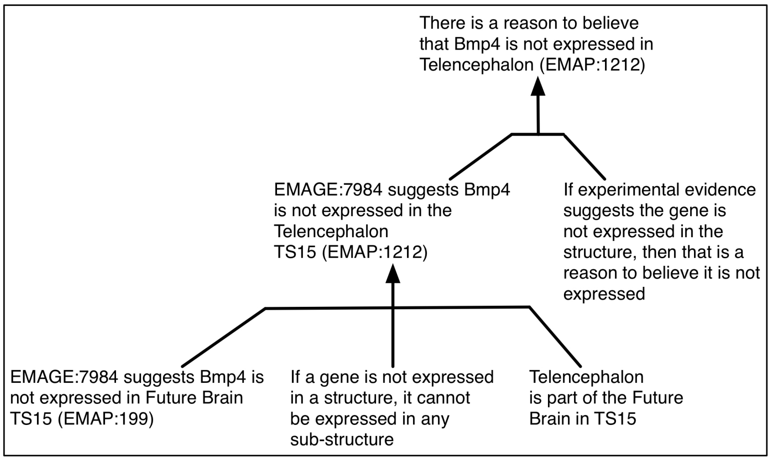

There is a reason to believe that Bmp4 is not expressed in Telencephalon (EMAP:1212)
EMAGE:7984 suggests Bmp4 is not expressed in the Telencephalon TS15 (EMAP:1212)
If experimental evidence suggests the gene is not expressed in the structure, then that is a reason to believe it is not expressed
EMAGE:7984 suggests Bmp4 is not expressed in Future Brain TS15 (EMAP:199)
If a gene is not expressed in a structure, it cannot be expressed in any sub-structure
Telencephalon is part of the Future Brain in TS15

Figure 7.11: Example of a standard tree-based argument visualisation, where the conclusion is at the top.

In this meeting the following key issues came to light. Firstly, the expert wanted to see all the information at once, and all in English. He rejected all the example presentation styles shown to him, and instead volunteered the idea of presenting arguments as a natural language paragraph. Secondly, the expert suggested there was a need to provide a visual summary of the arguments that quickly allowed the user to determine which arguments supported the notion of the gene being expressed and which took the opposite view. However, he did not suggest an implementation nor provide any guidance that could lead to an implementation. Thirdly, the expert believed there needed to be a wider range of experimental types included in the system (only *in situ* hybridisation was used in prototype one). In particular, he suggested trying to include SAGE data from the CGAP database.

The knowledge gathered during this meeting is implemented in the second prototype using the schemes and confidence values finalised after the sixth scheme creation meeting.

7.6 Prototype two

The second prototype was created during the Sealife project (see Section 8.1), with two purposes in mind. Firstly it had to represent the evolution discussed with the expert following the first prototype, and secondly it had to be used with an evaluation. The evaluation will be discussed further in Section 7.7 before being detailed in Chapter 8.

7.6.1 Outline of prototype

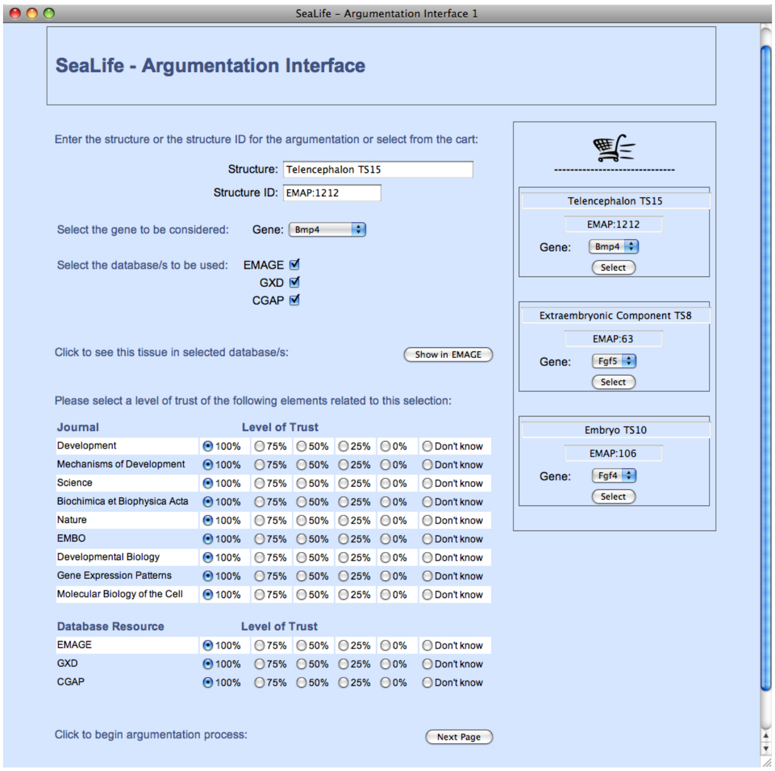
The minor differences between prototype one and two, in terms of the architecture and order of processes, are discussed in Section 7.3.

This prototype is developed in two separate parts using the finalised schemes and confidence values. The first part is the argumentation system, which is a subtle evolution of prototype one. The second part is a new graphical user interface (GUI).

The argumentation system uses amended schemes and strengths (prototype one being built before the schemes were finalised and the strengths assigned). It employs a different method of presenting arguments - adopting the expert's suggestion of a natural language paragraph. Otherwise it is similar, but not identical, to the original prototype. The system works, but has never been connected to its front end. This is because there is no need to do so in order to carry out the desired evaluation (see Chapter 8). Additionally, using pre-canned queries, data, and arguments from the argumentation system, rather than a true connection to the argumentation system, ensures all evaluation subjects receive an equal experience during the evaluation.

7.6.2 Walk through

The second prototype has a very similar workflow to the first. Initially (e.g. Figure 7.12) the user specifies the gene-tissue pair (s)he is interested in, and then indicates his (her) level of confidence in various journals and resources. On the second page, once the biological information has been extracted from the resources, the user is asked for his (her) confidence in the relevant researchers. The final page (e.g. Figure 7.13) provides an initial textual summary (the gene is (not) expressed), followed by a visual summary of the various arguments, and finally the arguments themselves.



SeaLife - Argumentation Interface

Enter the structure or the structure ID for the argumentation or select from the cart:

Structure: Telencephalon TS15

Structure ID: EMAP:1212

Select the gene to be considered: Gene: Bmp4

Select the database/s to be used: EMAGE ☑ GXD ☑ CGAP ☑

Click to see this tissue in selected database/s:

Please select a level of trust of the following elements related to this selection:

Journal	Level of Trust				
Development	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Mechanisms of Development	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Science	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Biochimica et Biophysica Acta	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Nature	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
EMBO	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Developmental Biology	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Gene Expression Patterns	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
Molecular Biology of the Cell	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know

Database Resource	Level of Trust					
EMAGE	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
GXD	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know
CGAP	● 100%	○ 75%	○ 50%	○ 25%	○ 0%	○ Don't know

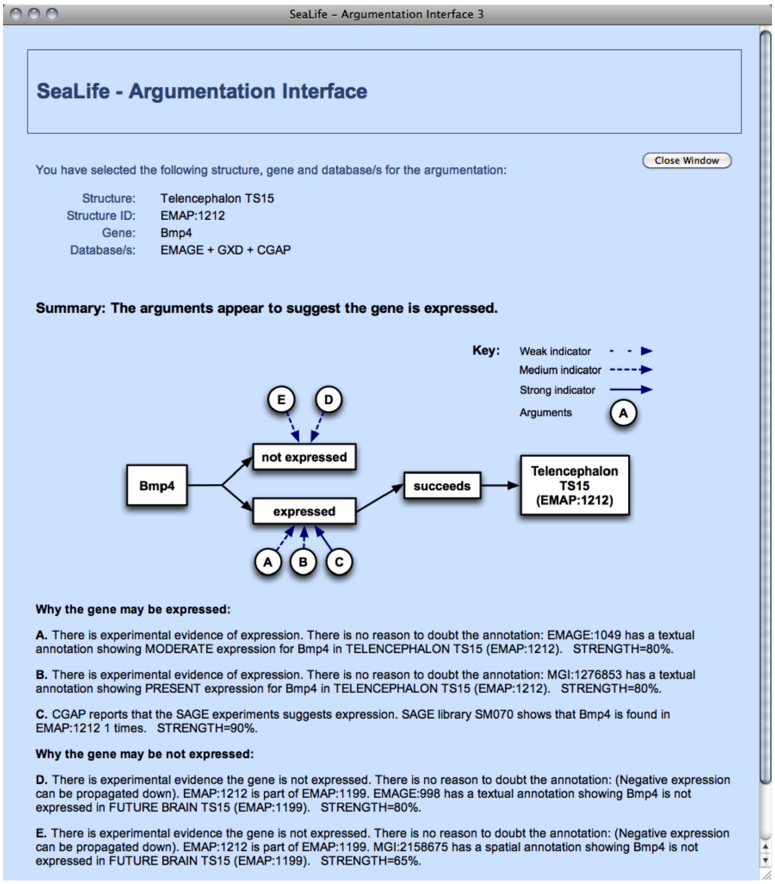
Click to begin argumentation process:

Telencephalon TS15
EMAP:1212
Gene: Bmp4

Extraembryonic Component TS8
EMAP:63
Gene: Fgf5

Embryo TS10
EMAP:106
Gene: Fgf4

Figure 7.12: First page of second prototype: asks the user which gene-tissue pair to consider and a range of confidence questions.



SeaLife - Argumentation Interface

You have selected the following structure, gene and database/s for the argumentation:

Structure:	Telencephalon TS15
Structure ID:	EMAP:1212
Gene:	Bmp4
Database/s:	EMAGE + GXD + CGAP

Summary: The arguments appear to suggest the gene is expressed.

Why the gene may be expressed:

A. There is experimental evidence of expression. There is no reason to doubt the annotation: EMAGE:1049 has a textual annotation showing MODERATE expression for Bmp4 in TELENCEPHALON TS15 (EMAP:1212). STRENGTH=80%.

B. There is experimental evidence of expression. There is no reason to doubt the annotation: MGI:1276853 has a textual annotation showing PRESENT expression for Bmp4 in TELENCEPHALON TS15 (EMAP:1212). STRENGTH=80%.

C. CGAP reports that the SAGE experiments suggests expression. SAGE library SM070 shows that Bmp4 is found in EMAP:1212 1 times. STRENGTH=90%.

Why the gene may be not expressed:

D. There is experimental evidence the gene is not expressed. There is no reason to doubt the annotation: (Negative expression can be propagated down). EMAP:1212 is part of EMAP:1199. EMAGE:998 has a textual annotation showing Bmp4 is not expressed in FUTURE BRAIN TS15 (EMAP:1199). STRENGTH=80%.

E. There is experimental evidence the gene is not expressed. There is no reason to doubt the annotation: (Negative expression can be propagated down). EMAP:1212 is part of EMAP:1199. MGI:2158675 has a spatial annotation showing Bmp4 is not expressed in FUTURE BRAIN TS15 (EMAP:1199). STRENGTH=65%.

Figure 7.13: Results page of the second prototype, featuring a visual summary of the textual arguments shown further down the image.

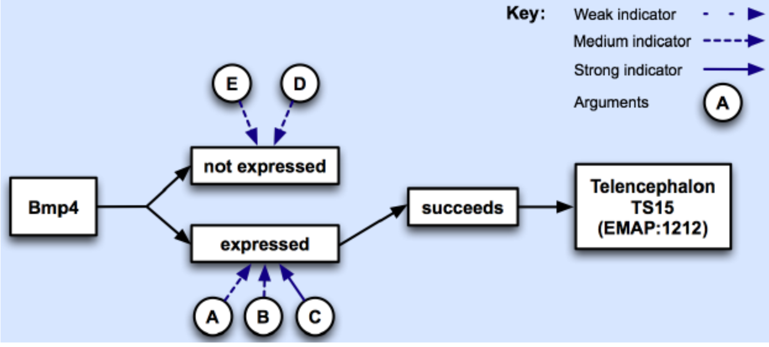


Figure 7.14: A close up of the argument summary from Figure 7.13.

The arguments are provided in the natural language paragraph style the expert stipulated. This is facilitated by the creation of a custom output generator for ASPIC[14]. Following each argument is the strength (expert's confidence) in the argument,

[14] Developed with assistance from ASPIC developer Matt South.

which is automatically calculated by the argumentation engine from the constituent sub-arguments using the weakest link method.

The visual summary (see Figure 7.14) provides an indication of which arguments are stronger (expressed or not expressed) and tells the user which side each argument supports. On the left side of the summary is a box containing the symbol of the gene, and opposite is a box containing the tissue. The complete path between the two indicates which is stronger, expressed or not expressed. Arguments themselves are circles, containing letters referencing individual arguments. Arrows link arguments to the conclusion they support. The style of the arrow indicates the strength of the argument.

In ASPIC the strongest argument wins the debate. Therefore, the strongest argument decides whether or not a gene is expressed - it is labelled a *strong indicator*. Arguments with a strength of less than 50% are deemed weak. Everything else is a *medium indicator*. The strongest argument dictates the content of the initial textual summary.

7.6.3 Limitations

Due to timing issues, the arguments provided for prototype two's evaluation use a limited range of schemes[15]. In particular, only the "reasons to doubt" are implemented fully. This reduces the total number of arguments as there are none arguing for the accuracy of the annotations. This is not a problem because the goal of the evaluation is not to verify the performance of the system. Instead, the aim is to determine whether or not the users can understand and interpret the information presented to them. There is no need to verify the system's accuracy because the goal is not to make a decision, merely to provide information that can be employed to help the users do so. It is worth noting that the evaluation indirectly comments on the accuracy of the system by reviewing the schemes.

A further issue is that the rules are aggregated. For each reason to doubt an annotation a corresponding inference rule is generated. A second set of rules collects all the reasons to doubt, and groups all the active reasons under one new conclusion.

[15] A final list of rules appears in Appendix B, this is not the list of rules used in this prototype, but a full list compiled after the evaluation when more time was available for this work.

Regardless of the number of reasons, there is only ever one of the second type of conclusion. The same is true for the reasons to doubt an experiment. This approach has two effects. Firstly it reduces the number of available arguments, and secondly it reduces the strength of the remaining argument to the lowest denominator. For example, assume there are three reasons to doubt an experiment with expert assigned confidences 15%, 25% and 45%. There should be three arguments, each with a different strength. In reality, there is one argument, which has the confidence 15% because of the weakest link principle.

Arguments in ASPIC are settled according to the strength of the arguments, resultantly this could affect the outcome in some situations (depending on the underlying experimental data). Fortunately this does not affect the examples used for the evaluation as the strongest arguments come from annotations for which there are no reasons to doubt the annotation or the underlying experiment. Hence there is no aggregation.

The final limitation relates to the inclusion of CGAP data. The expert wished this to be added, and it was felt to be a good idea because it would emphasise that the system was more than just an amalgamation of EMAGE and GXD. However no expert was available for this resource. Instead the original expert provided a basic level of knowledge, which would require to be extended in a real system. Accordingly, the inclusion of CGAP is not dwelt upon in this document.

7.7 Questions for evaluation

The second prototype makes a number of assumptions based on the expert's feedback from the first system, and such assumptions must be tested. Furthermore there is a need to assess the accuracy of the schemes (and associated strengths) developed earlier in this work. For these reasons an evaluation was performed: it shall be discussed in the next chapter. This section will concentrate on expressing the assumptions to be tested.

The foremost assumption is that the expert is representative of his community. It is necessary to test this, even if only indirectly. For example, do other users agree with the need for a visual summary of the arguments? The initial assumption is also tested by querying users to discover if they all prefer textual argument presentations rather than visual ones in the way the expert did.

Secondly, it is believed that the textual arguments contain all the necessary information, and should satisfy biologists. This is based on the fact that the expert approved them. Is this belief valid? Can biological users interpret the arguments and gain the information they need?

This issue of the user's ability to interpret the arguments also applies to the visual summary. The expert wanted one, but did not specify how it should appear. Consequently, the images are conjecture based on previous work with a range of biological users. Therefore it needs tested to ensure it is understandable.

7.8 Summary

During this chapter two prototypes are discussed - the second being an evolution of the first following expert feedback. The rationale for the creation of these prototypes, and the decisions behind their implementation are recalled in this chapter before they are both illustrated using screenshots. The goal of these systems was not to produce perfect tools, but to investigate argumentation.

Ultimately, it is the second prototype that is carried on through to the next chapter, Chapter 8, where an evaluation of one use of argumentation in biology is undertaken.

Chapter 8

Evaluation

This chapter reports on two of the evaluations undertaken during this work. The first is an exploration of the GUI and presentation mechanisms developed during Chapter 7. The second is an extension, focusing on one particular issue highlighted during the first evaluation.

Section 8.1 describes the motivation behind the first evaluation, the process undertaken to perform it, and the results obtained. Section 8.2 concentrates on the second evaluation, which is an online survey investigating the best visual presentation of arguments.

Although this chapter presents the results of the evaluations, a discussion and analysis of those results is reserved until Chapter 9.

8.1 Argumentation system evaluation

Chapter 7 discusses the creation of two prototype systems that use argumentation to resolve inconsistencies between EMAGE and GXD. It is the second prototype that is evaluated here - of particular interest is the user interface and the users' interaction with the arguments.

This work was performed during the Sealife project[1]. Sealife was an EU FP6 project examining the creation of a semantic web browser for the life sciences. As part of the contribution from Heriot-Watt University two systems were developed: this system, and GGAPS [158]. GGAPS assisted biologists and bioinformaticians in

[1] www.macs.hw.ac.uk/bisel/sealife

planning and executing bioinformatics workflows online. The system took a piece of bioinformatics information (for example a gene symbol) and determined the online services that the information could be submitted to. The user selected a workflow and the system automatically executed it.

Both systems were evaluated at the same time; though, details of the GGAPS system will be minimised in this document.

It is important to note the contribution of Dr. Gus Ferguson to this evaluation. As an employee of the Sealife project, in his role as evaluation expert, he supervised and documented much of this evaluation. He advised on the best methods to gather the information desired from the evaluation, acted as the observer during the evaluation sessions, and analysed the results before editing the documentation of this work [159][2]. The rest of the work remained the responsibility of the author, including: the decision of what to test for; the creation of the scenarios; the role of lead evaluator; the interpretation of the results; and, the writing of some sections of the final report.

8.1.1 Aim of evaluation

The evaluation aimed to achieve a number of goals. Standard methods and protocols were used to evaluate usability and obtain feedback on the prototype GUI. Bespoke protocols were used to obtain a summative and formative evaluation of the functionality. This included additional evaluation for specific aspects of the conceptual design of data presentation and visualisation. Moreover the evaluation was used as a focus for expert user evaluation of the underlying system processes by select users.

In particular, this evaluation was used to explore the following questions from Chapter 4:

- How should argumentation be presented to a biologist so that (s)he can understand, and utilise it?
- Is a graphical form of argument presentation better than a textual form?
- How can all of the (users') preferences be combined into a single interface?

In addition, Section 7.7 proposed a number of issues to consider:

[2] That document forms the basis of the current chapter.

- After basing much of the system's design on the opinion of an expert, does that expert truly represent his community?
- Are the textual arguments complete? Do they contain all the necessary information for a biologist to reach a decision?
- Are the users able to interpret the textual arguments, and the visual summary of the argumentation?
- Are the schemes accurate? Are the associated strengths appropriate?

Remember that the decision as to whether or not a gene is expressed will be controversial. Therefore, the system was not trying to provide a "correct" answer as no such answer may exist. Consequently, the evaluation did not concentrate on investigating the accuracy of the argumentation. Instead, it tried to focus on the usefulness of argumentation, and the system developed in Chapter 7, for assisting users to make a decision by themselves.

8.1.2 Method

User group and users

A user group of 18 people was evaluated. Of these 10 worked on EMAP in various roles, ranging from system developers to biological database curators. From this group were drawn the expert users whose function was to give expert opinion on aspects of the system. The other 8 users were postgraduate students at Heriot-Watt University.

The 18 strong user group comprised people with varied academic and training backgrounds. All users had degrees, the breakdown of which can be seen in Table 8.1.

Subject	**No. of users with**			
	BSc. degrees	**MSc. degrees**	**PhDs.**	**Research/ industrial experience**
Computing / IT	8	7	4	5
Biology / Genetics	5	0	3	2
Bioinformatics	0	3	2	4
Biochemistry	1	0	0	0
Psychology	2	1	0	1
Exercise / sports science	1	0	0	1
Geophysics	1	0	0	0

Table 8.1: Background of the users in the evaluation.

14 of the users were male. English was the first language for 7 of the 18 users, although all users spoke and read English well.

For the analysis of the results, the users were divided into two groups:

1. Half of the users were classified as biologists (referred to as biologist group), with formal qualifications and training in biology. 7 of these users were from EMAP, and 2 were postgraduate students.

2. The remaining 9 users were classified as non-biologists (referred to as non-biologist group), with little or no formal biological background. 6 were postgraduate students from the computer sciences, life sciences and psychology departments, and 3 were from EMAP.

Occasionally, the *biologist group* was subdivided into those who were expert in *in situ* gene expression (4 users) and those whose area of biological expertise was elsewhere (5 users).

The size[3] of the user groups meant that there could be no statistically significant

[3] For statistically meaningful differences to occur between user groups, each group must contain at least 30 users.

differences between them; however, on some occasions minor differences did occur, and if deemed noteworthy, they shall be highlighted.

Procedure

Test scenarios were created based on typical tasks the system was intended to perform. The GUI was then hard-coded with data taken from the databases, and arguments generated by the system developed in Chapter 7.6. This forestalled issues of availability and variation of third-party online resources during the course of the evaluation. Appropriate time delays were introduced to reflect delays experienced when the system accesses online resources.

Users were asked to complete a consent form which included a brief explanation of the systems and their purposes. They then filled in a questionnaire on their background, training and familiarity with appropriate bioinformatics online resources and journals. Subsequently the users completed the GGAPS scenario and associated questionnaire, before the argumentation scenario and questionnaire, and finally a general system usability questionnaire.

The evaluation was conducted using an Apple Macintosh computer (running the Firefox browser) in a dedicated room. Two observers were used for the evaluation, one interacting directly with the user, and the other observing the user and their interaction with the system, recording timings, errors, comments made by user and general observations on user actions. A script was used for consistency of order and procedure and users were prompted to comment, ask questions or ask for help at any stage during the evaluation.

The first argumentation scenario consisted of a walkthrough relating to the expression of a gene in the developmental mouse brain, using the default settings. The users were presented with the system's output (e.g. Figure 7.13) consisting of a textual statement of what the strongest argument indicated, followed by a visual summary of the arguments, and finally textual representations of the generated arguments. Users were asked specific questions regarding the process and presentation of the results. The second scenario involved walking through the same example, but altering selections for some of the parameters. The users were then presented with the argumentation modified appropriately to the altered parameters, and again asked specific questions regarding the results and their presentation. Finally specific questions regarding the users' understanding of, and views on, argumentation and the process of the system were posed. More details of the scenarios can be found in the argumentation scenarios and questionnaire protocol in Appendix F.

Protocols

The protocols used for the evaluations are attached in Appendix F, they include:

- Script
- Time-error form
- Consent form
- Background questionnaire
- Argumentation scenarios and questionnaire
- Usability questionnaire

Please note that the GGAPS centric protocols are not included.

The general usability questions were adapted from Shneiderman's *questionnaire for user interaction satisfaction* (QUIS) [160]. The QUIS format was used where appropriate for the other questionnaires along with more specific question formats, such as *Likert scales* [161].

The protocols were drawn up and refined following iterative piloting with a small number of postgraduate students.

8.1.3 Results: evaluation of argumentation system

This chapter will document the responses received without listing the exact questions that were posed. Thus, in order to truly understand the following text it is recommended to consider it alongside the protocols contained in the Appendices.

Background knowledge and familiarity with bioinformatics/molecular biology

Users were asked to rate their familiarity with six bioinformatics tools and databases used in the evaluation on a scale of 1 to 9 representing "Never Used" to "Very Familiar". The results (see Table 8.2) show that the non-biologist group was generally unfamiliar with all the databases and tools, while the biologist group was generally familiar with them; however, the biologists were unfamiliar with a small number of resources including the SAGE gene expression resource CGAP.

	Non-biologist group	**Biologist group**
	Median	Median
XSPAN	1	1
EMAGE	1	4
GXD	1	6
BLAST	1	3
UniProt	1	3
CGAP	1	1

Table 8.2: How familiar the users are with a range of resources on a scale of 1 - 9, with 9 representing very familiar, and 1 never used.

Users were asked to rate their knowledge from 1 = No Knowledge to 9 = Expert on the following:

Gene Expression

Biologist group ranging from 4 to 9 with a median of 6 (mean = 6.0);

Non-biologist group ranging from 1 to 4 with a median of 2 (mean = 2.3).

Ontologies

Biologist group ranging from 1 to 8 with a median of 5 (mean = 4.7);

Non-biologist group ranging from 1 to 7 with a median of 2 (mean = 3.1).

Users were asked to indicate their familiarity with nine commonly used biological and bioinformatics journals using the following options:

- read all of journal;
- read some papers;
- know journal but do not read it;
- never heard of journal.

The results are summarised in Table 8.3.

The background information gathered on users was particularly useful for adding context to comments users made during evaluation, and was the basis for the division of the biologist group into *expert* and *other biologists*.

	All (N = 18)			
	Read all	**Read some**	**Know, do not read**	**Never heard of**
Development	0	5	4	9
Mechanisms of development	1	9	6	2
Science	1	9	6	2
Biochimica et Biophysica Acta	0	1	6	11
Nature	0	10	5	3
EMBO	0	2	5	11
Developmental Biology	0	4	5	9
Gene Expression Patterns	0	3	3	12
Molecular Biology of the Cell	0	4	5	9

Table 8.3: User familiarity with common biological and bioinformatics journals.

Time taken to complete argumentation scenarios

Table 8.4 lists the times taken by each group to complete the argumentation scenarios.

	Non-biologist group	**Biologist group**	**All** (N = 18)
Argumentation Section A			
Mean	4 m 22 s	4 m 45s	4 m 34 s
Minimum	2 m 56 s	2 m 20 s	2 m 20 s
Maximum	7 m 00 s	10 m 20 s	10 m 20 s
Standard deviation	1 m 15 s	2 m 33 s	1 m 57 s
Argumentation Section B			
Mean	2 m 36 s	3 m 31 s	3 m 04 s
Minimum	0 m 54 s	1 m 49 s	0 m 54 s
Maximum	4 m 17 s	4 m 42 s	4 m 42 s
Standard deviation	1 m 08 s	0 m 56 s	1 m 06 s

Table 8.4: Duration of evaluation.

Journals and authors

During the creation of the schemes (documented in Chapter 6) the expert advised that the users' confidence in the journal and the research team might affect their confidence in the experiment (and thus the result). For this to be true, the user must be able to recognise both the journal and author(s). Consequently, users were asked to indicate their confidence in the journals commonly used to publish gene expression experiments by ranking them from 1 (highest confidence) to 9 (lowest confidence), plus a "Don't Know" option. Table 8.5 shows the results for the rankings (excluding "Don't Know"): there appears to be little difference between the journals. Most of the rankings were provided by the biologist group.

Journal	**Rankings**			
	No. of rankings	**Mean**	**Range**	**Standard deviation**
Development	10	2.7	1 - 6	1.6
Mechanisms of development	6	2.8	1 - 7	2.2
Science	14	3.4	1 - 7	2.4
Biochimica et Biophysica Acta	6	4.2	1 - 8	2.6
Nature	13	3.3	1 - 7	2.3
EMBO	8	3.0	1 - 7	2.3
Developmental Biology	10	2.7	1 - 7	2.1
Gene Expression Patterns	9	2.0	1 - 5	1.2
Molecular Biology of the Cell	8	2.3	1 - 4	0.9

Table 8.5: User confidence in journals - very high (1) to very low (9).

Users were asked to indicate if they recognised authors who publish in the field of *in situ* gene expression for the developmental mouse - results are in Table 8.6: most authors were only recognised by 2 subjects (half of the expert group). All of the responses came from the biologists.

Author	No. of users recognising
Furuta Y	3
Hogan BL	4
Trainor PA	2
Hebert JM	2
Martin GR	2
Niswander L	2

Table 8.6: User recognition of authors.

Scenario one

Scenario one involved interpreting the output of the system, when the trust parameters were left at their defaults.

When asked what the result was:

- 15 users (83%) correctly identified the result;
- 3 users (17%) responded “Don’t Know”.

When asked to identify the strongest argument in the output:

- 13 users (72%) correctly identified the strongest argument;
- 2 users (11%) identified all three supporting arguments as the strongest argument;
- 3 users (17%) did not answer.

Users based their identification of the result on a range of things, including:

- on the strongest argument (3 users);
- all three positive arguments;
- three supporting against two negative arguments;
- weight of arguments in favour;
- expert knowledge of experiments (implies ignored system results);

- details of experiment;
- conclusion of system (2 users);
- graphical presentation of arguments (2 users);
- summary data and strengths quoted in textual details of arguments;
- no decision (3 users).

Scenario two

The second scenario used the same example, and asked the same questions, but this time the trust parameters were set to prescribed levels.
When asked what the result was:

- 12 users (67%) correctly identified the result of the argumentation process;
- 3 users (17%) answered incorrectly;
- 3 users (17%) as"Don't Know".

When asked to identify the strongest argument in the output of the second scenario:

- 11 users (61%) correctly identified both of the two equally strong arguments;
- 4 users (22%) identified only 1 of the arguments;
- 1 user (6%) said: "C or D";
- 2 users (11%) merely quoted the argumentation result.

Again, the users' identification of the result was based on a variety of elements:

- on the 2 strongest arguments (7 users);
- summary data and strengths quoted in textual details of arguments (4 users);
- graphical presentation of arguments;
- no specific answer - made comments on task (6 users).

User opinion of system

Directly following the second scenario, the subjects were asked to fill in a questionnaire designed to gather their opinions on a number of areas of the system including:

- The questions on trust (journals and authors);
- The amount of information contained in the arguments;
- The ease of understanding the textual arguments;
- The usefulness of the graphical summary of the arguments (e.g. Figure 7.14).

The results are summarised in Figures 8.1 to 8.5. In general, those results indicated that the users were happy with the system, approving all aspects. In most results the differences between the user groups were not significant; though, when asked how easy it was to understand the textual arguments (Figure 8.4) the non-biologist's ratings were significantly lower with a median of 3 compared to the biologist's median of 7 ($p = 0.0121$).

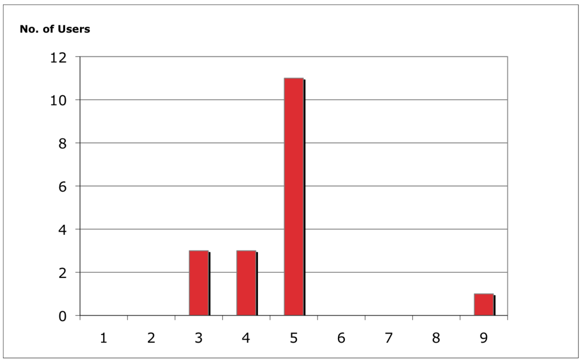


Figure 8.1: Were the questions on trust of journals and researchers asked of you too few (1) - too many (9)? Median = 5.

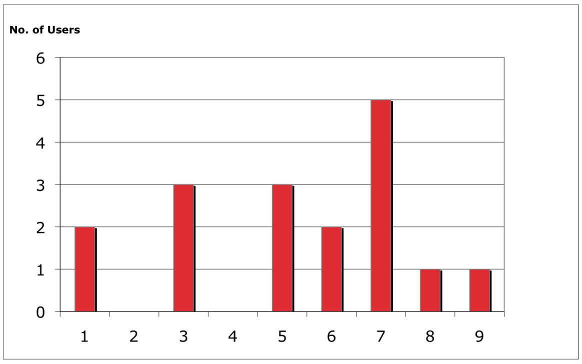


Figure 8.2: Would the option to hide the questions on trust of journals and authors be, undesirable (1) - very desirable (9)? Median = 6.

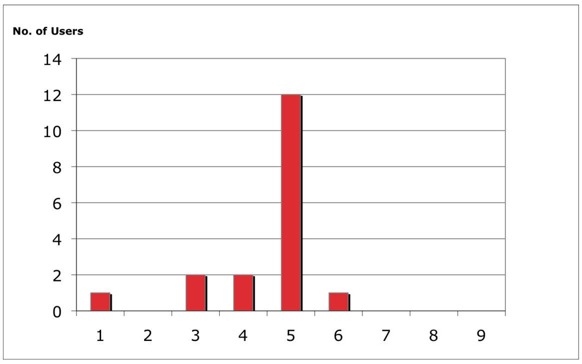


Figure 8.3: Was the amount of information presented in the arguments, too much (1) - too little (9)? Median = 5.

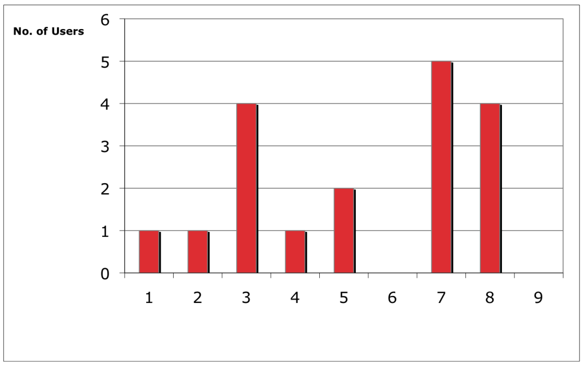


Figure 8.4: How well did you understand the information presented in the arguments, not at all (1) - completely (9)? Median = 6.

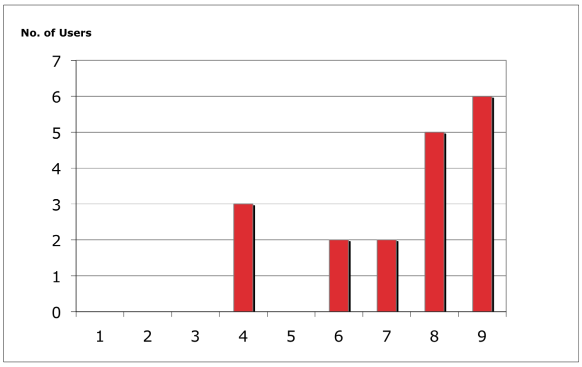


Figure 8.5: How helpful were the diagrams of the arguments, totally unhelpful (1) - very helpful (9)? Median = 8.

Presentation and perception of textual arguments

Users were given a typical argument generated by the argumentation system and asked what they inferred from this argument. A perfect answer stated the gene was not expressed, mentioned the notion of propagation and acknowledged that the argument was defeasible. A correct answer mentioned two of the three components of a perfect answer. The results broke down as follows:

Expert biologists

- 1 gave a perfect answer;
- 1 gave a correct answer

- 1 restated the conclusion;
- 1 could not be classified.

Other biologists

- 1 gave a correct answer then introduced an error;
- 2 restated the conclusion;
- 2 were not classifiable.

Non-biologists

- 3 restated the conclusion;
- 6 could not be classified.

10 users (56%) correctly inferred that the argument supported the gene not being expressed in the tissue. However, of these 10 users, 2 (11%) gave erroneous interpretation of the STRENGTH=79% statement shown below (direct quotes):

- 4:1 people who've researched the area believe this to be the case (strength=79%);
- this finding has emerged from a trusted source (100%);

Comments from users who had produced erroneous remarks included (direct quotes):

- although argument strong, don't trust underlying data;
- the brain development is abnormal;
- need special knowledge to understand the language strength;
- 30% understandable for me;
- too confusing.

Users were asked how this textual method of presentation could be improved. 11 users thought it did not need improvement, or did not comment. The comments of the others are summarised below:

- graphical representation needed (2 users);

- more structure to the text, e.g. bullet points (2 users);
- more prominence for expression level (2 users);
- more explanation/definition of different parts e.g. "STRENGTH = 79%" (2 users).

Presentation and perception of graphical arguments

Users were then presented with a graphical representation of a typical argument. 12 users correctly inferred that the gene was not expressed in the tissue. 2 users gave comments which did not indicate their inference ("did not understand how the arguments connect" and "trusted sources, but not much else, this is a little more confusing than the diagram presented on screen"[4]). 4 users did not answer the question.

Users were asked to suggest how the graphical presentation could be improved. The results of the 6 responses are summarised below:

- drawing the graph top-down instead of bottom-up (3 users);
- like to see strength information;
- presented as it was on the computer[3];
- the diagram suggests that there is more than one answer when there is only one, a paragraph or more fluent flow diagram may be better.

In response to the first comment, a follow-up study was undertaken to evaluate users responses to different graphical representations of arguments - the results are presented in Section 8.2.

Comparison of graphical and textual argument presentations

Further specific questions on both the textual and graphical presentations of the argument were put to users. The results are given in Figures 8.6 to 8.9.

[4] It was assumed that this referred to the graph displayed in the scenarios; however, that graph summarised the arguments rather than depicting a single argument.

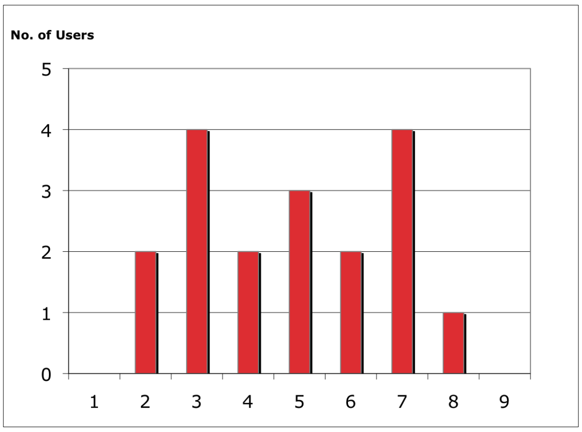


Figure 8.6: How easy is this textual argument to understand, very confusing (1) - very straight-forward (9)? Median = 5.

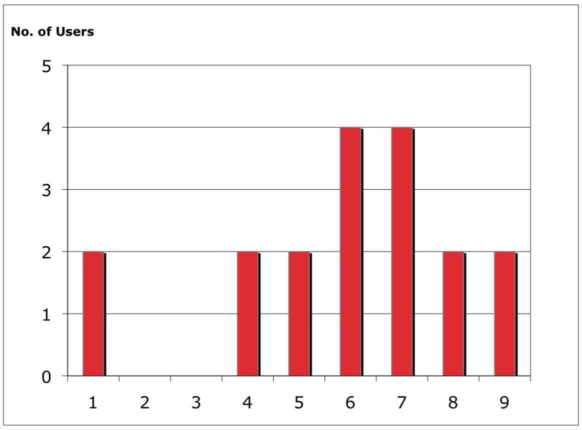


Figure 8.7: How easy is this graphical argument to understand, very confusing (1) - very straight-forward (9)? Median = 6.

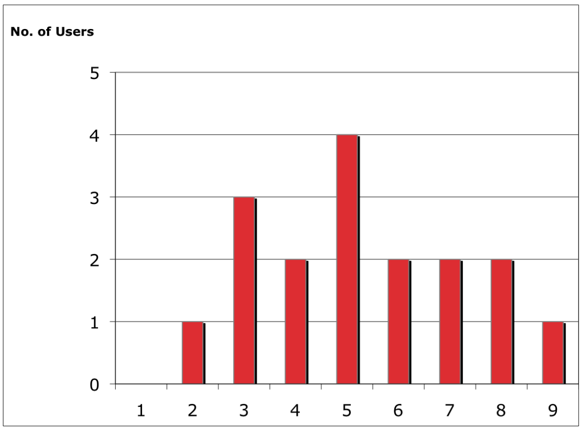


Figure 8.8: How intuitive do you find the inference drawn in the textual argument, not intuitive (1) - very intuitive (9)? Median = 5.

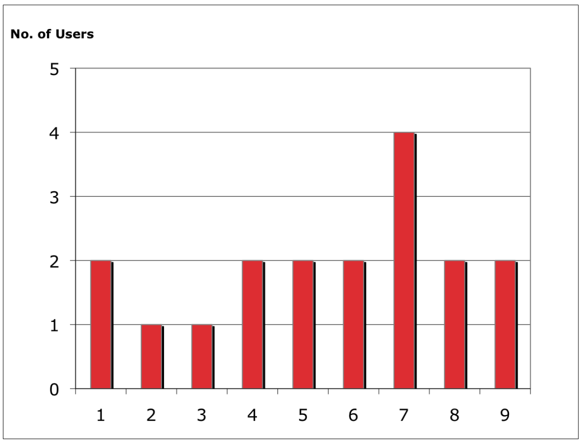


Figure 8.9: How intuitive do you find the inference drawn in the graphical argument, not intuitive (1) - very intuitive (9)? Median = 6.

When asked which form of presentation users favoured, 8 users preferred the graphical presentation, 8 chose the textual presentation, 1 answered both (this was treated as a "Don't Know"), and 1 did not know which they preferred. The responses of the user groups are compared in Table 8.7 which shows the expert biologists were split equally between preferring the textual and graphical presentations of the arguments, while the two other groups were divided almost evenly.

Preferred	**Expert biologists** (N=4)	**Other biologists** (N=5)	**Non-biologists group** (N=9)
Textual	2	2	4
Graphical	2	3	3
Don't know	0	0	2

Table 8.7: User preference of argument representation.

Visual argument summary

A visual summary, such as Figure 7.14 from Chapter 7, provides a synopsis of the outcome of argumentation process. Figure 8.5[5] considers the users' opinion of the visual summary of the arguments produced by the system; however, the motivation was to ensure that users could interpret the graph accurately.

Initially users were asked to say whether the diagram suggested that the gene was expressed or not. 10 users said that the gene was not expressed (correct conclusion), 2 said the gene was expressed (incorrect) and 6 selected "Don't know". This equated to 44% of users being unable to use the diagram to reach the correct conclusion.

Users were then asked to judge from the diagram what they thought the strengths of three of the arguments were. The results are shown in Table 8.8.

[5] Although the wording suggests the users were being asked to judge diagrams depicting arguments, at this stage the only diagram they had seen was the visual summary.

Argument	**User rating**			
	Strong	**Medium**	**Weak**	**Don't know**
Argument H (correct: STRONG)	16	1	0	1
Argument C (correct: WEAK)	0	0	16	2
Argument A (correct: MEDIUM)	0	15	1	2

Table 8.8: User judgment of argument strength.

Two users (both software developers from EMAP) were responsible for all the "Don't knows". The two incorrect answers came from different users. The remaining 47 answers (87%) were correct.

When asked to rate how easy the diagram was to understand from 1 (very easy) to 9 (very hard), the median was 2 - see Figure 8.10.

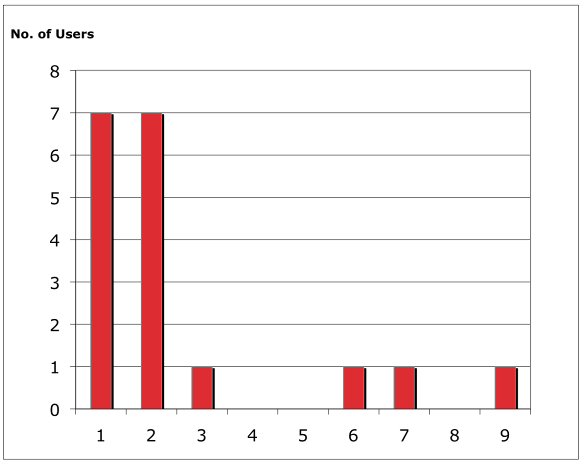


Figure 8.10: How easy was the argument summary diagram to understand, very easy (1) - very hard (9)? Median = 2.

6 users had no suggestions on how this method of presentation could be improved; the other suggestions are summarised below:

- need to improve distinguishing of strong/medium/weak arguments (5 users) with suggestions including:

  - coloured arrows (3 users);
  - colour coded arguments.

- interactive diagram with suggestions including:
  - be able to click/mouse-over on each argument to see details;
  - maybe you could toggle arguments on and off to see different results.
- filtering of results;
- explanation/more information of background process/studies leading to rating of argument strength (4 users);
- conclusion arranged closer to gene;
- need to know relationships between weak, medium, and strong.

System results page

Users were presented with the output from the system in the format used in the scenarios (one line natural language summary, summary diagram and natural language arguments). Users were asked which sections they would use to make their own decision. The results are summarised as follows:

- 6 users used all three elements;
- 3 used the summary and the diagram;
- 3 used the diagram only;
- 3 used the diagram and the textual arguments;
- 2 used the textual arguments only;
- 1 used the summary and the textual arguments;
- no users relied only on the summary.

Focusing solely on the expert users:

- 2 users applied the summary, 2 did not;
- 3 subjects used the diagram;
- 3 applied the textual arguments.

Additional comments

Users were asked for additional comments on the system and presentation of the argument process and these are summarised below (direct quotes):

Presentation of arguments

- Illustration is extremely useful for an overview, although text is useful to determine strength of the argument developed;
- Would like to have more explanation of argumentation process and why one argument wins over another;
- Would like to be able to view each argument as a graph.

Issues of detail

- I may look at detailed descriptions if I need to do more work (dry or wet) and would like to be able to navigate different levels of detail where desired;
- The diagram of arguments is helpful but the detailed description is needed as well. The diagram is just the simplified version (summary) of the text and I don't trust the diagram alone;
- Would use personal expertise of system to re-query based on argument and look at more detail e.g. link to database would be useful;
- Would like to be able to view some sort of personal profile for each researcher to help user determine level of trust;
- Should show on the results page how any changes of trust in journals /databases/ authors have altered the results and what changes in parameters have occurred.

The scenarios

- Not really comfortable with artificiality of example.

This is an abbreviated list of the comments, a full list can be seen in Appendix E.

8.1.4 Conclusions

Results provided earlier in this chapter divulge a number of insights into both the underlying argumentation based system, and the techniques used to present its output to biologists.

In the first scenario (default trust values) the majority of non-biologists correctly answered the question asking them to identify the strongest argument, in contrast only 1 expert and 3 other biologists answered correctly. In the second scenario (modified trust values), comments showed that the experts had problems accepting the scenario imposing changes in trust status for journals and authors. As they disagreed with the trust values, and felt the scenario to be unrealistic, they either did not answer the question or simply expressed their disagreement with the system, resulting in correct answers from less than half of the biologists' group. Again the non-biologist user group generally answered correctly.

Issues of trust in expert or decision aid systems have been addressed in work done since the late 1980s. One example is Muir [162] who raised the issue of modelling and measuring users trust in such systems and identified a phenomenon of spreading distrust in a system from one area to another. In the evaluation of the argumentation system, responses, comments and observations indicated that users appreciated the imposed conditions and contrived nature of the evaluation scenarios and there was no spread of distrust in the system beyond this issue. Work on trust in decision aid systems has been focused on modelling and measuring trust in individual systems [163]. Currently there does not seem to be any generalised method of measuring trust akin to QUIS for measuring usability.

Users were satisfied with the amount of information presented by the system, with 15 scoring it in the 4-6 out of 9 range, and 12 of those rating it at 5 (just right). Ratings of how well users understood the arguments (1 = Not at all - 9 = Completely), showed that the non-biologist groups ratings, with a median of 3, were significantly lower compared with the biologists with a median of 7 (p = 0.0121). Comments and observations indicated that most of the issues concerning the biologist group related to the strength ratings while the low levels of understanding among the non-biologists were due to a lack of biological knowledge.

From a detailed examination of the responses and observations, it emerged that the presentation of results from the argumentation system was clear and enabled users, who used the system as an expert system, to reach valid conclusions. However, experts seemed to have issues with trusting the argumentation system and the results it produced, making their use of the system problematic.

When considering the arguments, and their presentation, the textual argument alone appeared to be too complex. This combined with the poorly understood strength rating system, suggested that textual representation would be unsuitable as a sole presentation method. Most users found the graphical version simpler and more intuitive - the majority of the users found the argument easy to understand (median 6) and intuitive (median 6) with no significant differences between the biologist and non-biologist groups. However, half the experts preferred the textual form, implying that both mechanisms should be available. Regardless of the version used, it appeared that the explanation of the argument (reason for the inference step) needed further clarification.

Concentrating on the argument strengths, in Section 6.2 the reasons behind using numerical values for documenting the strength of an argument were given. Although there was clear evidence that the strength of an argument was often misinterpreted by users, there was no indication that this was because of the use of numerical values. Instead, the more likely explanation was that the biological users were confused by the term "strength" interpreting it in its biological sense rather than its argumentation one. This would be resolved by a better choice of label, for example: argument score.

Generally the results page was acknowledged by the users as a success, with responses showing that all three elements of the page (summary statement, summary image, textual arguments) were widely employed. Despite its popularity, it is worth noting that 8 of the 18 users (44%) were unable to determine which conclusion the summary image presented. However, the majority of users were able to ascertain correctly the strength of the individual arguments. Nevertheless, the summary diagram requires further development.

Overall it was evident that there were several areas requiring further work and examination:

- although graphical representations of arguments were easier for users to understand, some experts preferred the textual form, and in both representations the explanation of the argument (reason for the inference) needed clarification;
- the summary image may have been improved by greater use of colour;
- the questions on researchers should have been optional, and hidden by default;
- a more reliable and robust method of creating the strength levels for the arguments confidences was required.

These bullet points contain a number of subtle issues. Some consideration of these matters is provided in Chapters 9 to 11. Additionally, Section 8.2 briefly explores which form of graph-based argument representation is most suitable for the life sciences. Full resolution of these bullet points will require a series of detailed investigations that are beyond the scope of this thesis.

Moreover, the evaluation raised questions concerning the schemes used in this work; however, this discussion shall be withheld until Chapter 11.

Despite this further work the evaluation showed that the system was able to convey its message to users from a broad spectrum of backgrounds, and with varied biological expertise. The results from the non-biologist group demonstrate that the presentation of the system's results was clear and understandable, if the user accepted the results on trust. Members of the biologist group had difficulty with this, and employed their own knowledge to evaluate the information rather than using the information presented by the system.

An analysis of the results and conclusions from this evaluation, and what they mean for argumentation within biology, is provided in Chapter 9.

8.2 Online evaluation

The second evaluation concentrated solely on one question: several main forms of graphic presentation (of arguments) are featured in tools and publications, which of these was favoured by biologists? This evaluation was conducted in response to the previous evaluation. There users were asked to comment on a graphical depiction of a typical argument in which the argument was presented as a tree-like graph with the conclusion at the top, e.g. Figure 8.11. A number of users commented that the graph was upside down.

Consequently, it was decided to compare that initial graph, with the same representation rotated 180°, so that the conclusion was at the bottom (e.g. Figure 8.12). The former was colloquially known as the *bottom-up* graph, as it must be read from the bottom-up, and the second graph *top-down*. Another common form of argument presentation, reading from left-to-right, was proposed by Toulmin. It was decided to include a simplified form of this presentation in the evaluation (e.g. Figure 8.13). Each graph featured the same argument, which was an extended version of the graph from the first evaluation (extended to include extra information in the explanation).

8.2.1 Evaluation of argument graphs

The evaluation of the representations was undertaken as an online survey, with a user group drawn from biologists and bioinformaticians, including some staff from EMAP and participants recruited by email invitation through the *Scottish Bioinformatics Forum*[6] mailing list. The survey used the three graphs shown in Figures 8.11 to 8.13. The first query required participants to select their preference between the bottom-up and the top-down versions of the tree-based argument graph. The second question required them to select their preference between their chosen tree-based representation and the Toulmin-like graph. The participants' choices for both stages were submitted through an online form.

This evaluation made no attempt to record the background of its subjects. The fact that they either worked for EMAP or chose to subscribe to a mailing list for biologists and bioinformaticians in Scotland suggested that they were probably based in Scotland and worked in a biology related discipline. However, it was impossible to say which discipline. Furthermore, no attempt was made to record the reason for the users' preference.

[6] `www.sbforum.org`

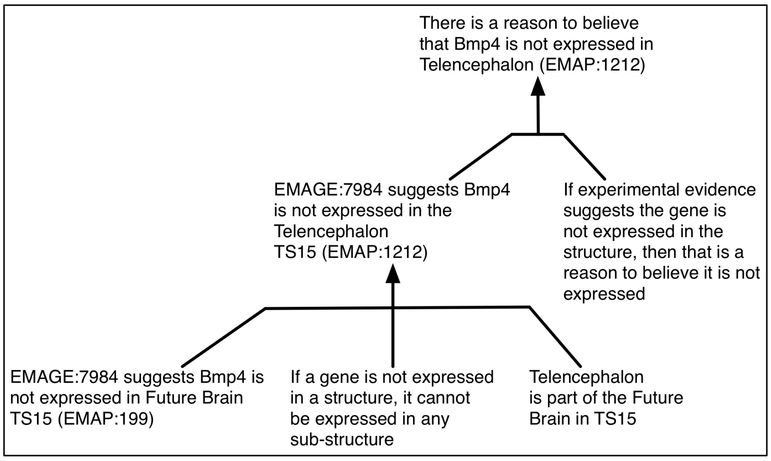

There is a reason to believe that Bmp4 is not expressed in Telencephalon (EMAP:1212)
EMAGE:7984 suggests Bmp4 is not expressed in the Telencephalon TS15 (EMAP:1212)
If experimental evidence suggests the gene is not expressed in the structure, then that is a reason to believe it is not expressed
EMAGE:7984 suggests Bmp4 is not expressed in Future Brain TS15 (EMAP:199)
If a gene is not expressed in a structure, it cannot be expressed in any sub-structure
Telencephalon is part of the Future Brain in TS15

Figure 8.11: Bottom-up version of argument.

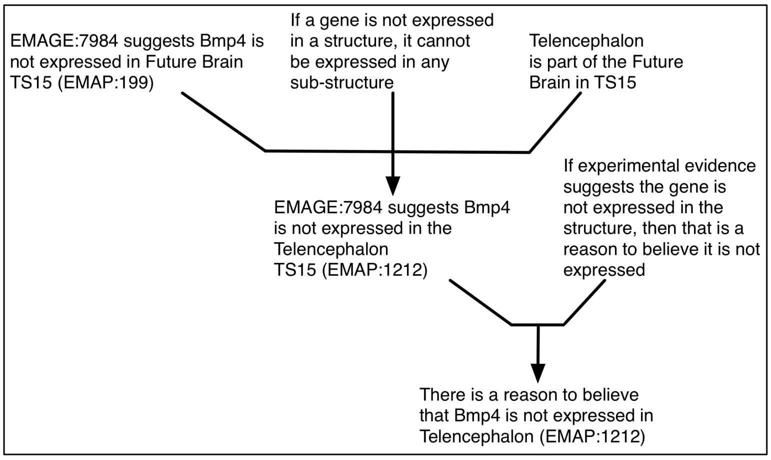

EMAGE:7984 suggests Bmp4 is not expressed in Future Brain TS15 (EMAP:199)
If a gene is not expressed in a structure, it cannot be expressed in any sub-structure
Telencephalon is part of the Future Brain in TS15
EMAGE:7984 suggests Bmp4 is not expressed in the Telencephalon TS15 (EMAP:1212)
If experimental evidence suggests the gene is not expressed in the structure, then that is a reason to believe it is not expressed
There is a reason to believe that Bmp4 is not expressed in Telencephalon (EMAP:1212)

Figure 8.12: Top-down version of argument.

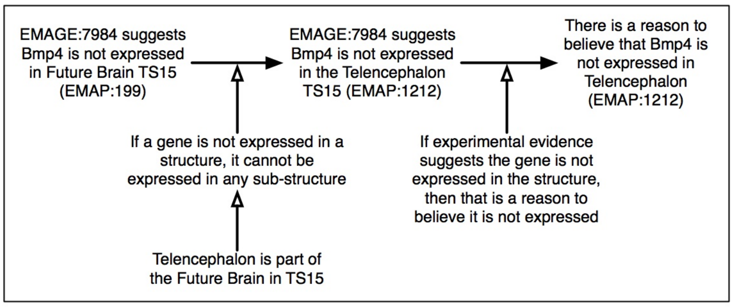

EMAGE:7984 suggests Bmp4 is not expressed in Future Brain TS15 (EMAP:199)
EMAGE:7984 suggests Bmp4 is not expressed in the Telencephalon TS15 (EMAP:1212)
There is a reason to believe that Bmp4 is not expressed in Telencephalon (EMAP:1212)
If a gene is not expressed in a structure, it cannot be expressed in any sub-structure
If experimental evidence suggests the gene is not expressed in the structure, then that is a reason to believe it is not expressed
Telencephalon is part of the Future Brain in TS15

Figure 8.13: Toulmin-like version of argument.

A total of thirty-eight participants responded. For the tree-type graph (bottom-up versus top-down) the top-down version was most popular with thirty-one respondents (82%) favouring it. In the second stage, the Toulmin-like graph was clearly favoured over the tree-type with twenty-four respondents (63%) indicating it was their preferred representation. Further analysis showed that twenty-three (74%) of the thirty-one who originally selected the top-down tree-graph chose the Toulmin-like representation at the second stage. In contrast only one (14%) of the seven users who initially favoured the bottom-up version choose the Toulmin-like graph as their overall preferred representation.

8.2.2 Conclusion

Although it was impossible to catalogue the participants in this study, it was legitimate to class them as biologists and bioinformaticians. In this case, the online survey clearly showed that individuals working in the biological community had a clear preference for the Toulmin-like graph. Furthermore, there was strong evidence that the default graphical representation of many argumentation theory scholars and systems (the bottom-up tree) was the least favoured depiction with this user group.

8.3 Summary

This chapter presents two evaluations of the second prototype discussed in Chapter 7.

The evaluations undertaken show that the basic GUI developed for the argumentation system was appropriate; however, it was not perfect. Important issues were raised that require attention before any similar system could be made publicly available.

In particular the second evaluation highlighted the importance of developing a range of visual presentation styles for arguments. Furthermore, it suggested that the standard presentation method used by default on many argumentation tools and toolkits is not appropriate for the target user group of this system. In addition, the first evaluation highlighted the importance of using both graphical and textual based arguments presentation styles.

This first evaluation raised doubts over the accuracy and relevance of some the schemes produced in Chapter 6 - something that will be commented on further in Chapter 11. Additionally, the assignment and presentation of argument strengths were shown to be inadequate. Clearly a more "natural" mechanism must be found. Despite these issues, feedback on the system was broadly positive, and indicated that such a system could prove useful.

An analysis of the output of this chapter, and its impact on argumentation within the biological world, is discussed next in Chapter 9.

Chapter 9

Analysis: presentation of arguments and argumentation

Earlier, in Section 4.6, a number of questions were raised for this thesis to deliberate upon. Further questions were posed in Section 7.7 whilst considering the prototypes constructed during Chapter 7. Chapters 9 to 11 shall provide answers to these questions by returning to work discussed in Chapters 5 to 8, and augmenting it with related work and analysis.

Initially the questions to be answered will be restated in Section 9.1.

9.1 Questions to be resolved: a summary

Chapter 4 motivates this work by exploring the inconsistency and incompleteness between, and within, two online resources publishing *in situ* hybridisation gene expression information for the developmental mouse. Argumentation is proposed as a mechanism to help end users resolve these issues. In doing so, it is acknowledged that a number of issues have to be tackled in order to demonstrate argumentation is a suitable approach. These challenges, stated as questions, are discussed in detail in Section 4.6. Here a reprise will be given.

Initially three high-level issues are documented:

1. Which form of argumentation is appropriate as a solution to the issues described in Sections 4.1 to 4.3 - and how effective is it?

2. How should argumentation be presented to a biologist so that (s)he can understand, and utilise it?

3. What insights have been gained from the work, and how does that inform the future use of argumentation within biology?

Question 1 will be explored in Chapter 10, and Chapter 11 discusses question 3. However, question 2 is related closely to the evaluation, therefore it will be considered in this chapter.

In Chapter 4, question 2 is extrapolated to provide the first three questions in Table 9.1. The remaining questions in Table 9.1 come from Section 7.7, where they are proposed in response to the prototype systems documented in Chapter 7.

Question	**Answered in**
Which is the best visual presentation mechanism for arguments?	Section 9.2
Which presentation mechanism - graphical or textual - seems most appropriate?	Section 9.3
How well do the standard argumentation concepts translate to the biological domain?	Section 9.4
Can a biological user interpret the information in the arguments correctly?	Section 9.5
Do the expert approved textual arguments provide all the information necessary for a biologist to reach a decision?	Section 9.6
Is the visual summary of the debate a worthwhile inclusion?	Section 9.7

Table 9.1: Summary of questions answered in Chapter 9.

It should be remarked that the presentation of arguments to a biologist is a complex matter. One obvious complication is that it is necessary to distinguish between an argument and argumentation. The former is a single justification, the latter a series of justifications situated inside a debate. Therefore to present argumentation it will be necessary to determine the best way of presenting arguments, and the context in which they are used. Creating a good solution, let alone the ideal one, will involve social scientists, psychologists, and experts from other disciplines working together.

Various evaluations and tests will need to be performed, and analysed. It is not feasible to do all of this here, and thus answer the overarching question. However, it is possible to produce a preliminary solution, evaluate that solution, and analyse it - Chapters 7 and 8 do this.

The subsequent sections (9.2 to 9.7) answer each of these questions from Table 9.1 in turn before a summary is provided in Section 9.9.

9.2 Which is the best visual presentation mechanism for arguments?

Traditionally argumentation tools/systems present arguments visually using the notion of a graph. Choosing the appropriate mechanism is important, as Scheuer *et al.* remark:

> ...the form of external argument representation (and accompanying interaction) does matter and, thus, should be seriously considered by system designers. The studies by Suthers and Hundhausen [164], Suthers et al. [165], Nussbaum et al. [166], McAlister et al. [167], Schwarz and Glassner [168], and Stegmann et al. [169] show that the way in which a system lays out an argument visually and allows students to use it has an impact on the behavior and learning gains of students. ([170] page 90)

Commonly, argumentation systems choose to visualise arguments as a bottom-up graph (conclusion at the top with facts and rules beneath), yet the comparisons of different styles of graph, in Section 8.2, suggest this is the least popular style with the biological user group. During the evaluation of the prototype system, see Section 8.1.3, a number of users commented that the bottom-up style seems to be upside down. This perception continues in the second evaluation, with 82% of users favouring the top-down graph over the bottom-up.

Overall, the most popular graph is the left-to-right graph loosely based on Toulmin's representation, which is in-keeping with the findings in other domains:

> Ye and Johnson [149] describe empirical studies demonstrating that argument structured explanations are the most effective in terms of bringing

> about changes in users attitudes toward rule-based systems. In particular, they indicate the significance of the Toulmin model for structuring explanations, in that identifying premises of rules as components of the Toulmin schema highlights discrete response steps that an explanation facility should follow in order to convincingly answer user queries. ([150] page 33)

A likely cause of this popularity is that it is natural to read from left-to-right. However, this presentation may not work so well in cultures where the norm is to read from right-to-left. Although the bottom-up graph is the least popular representation, it has roughly 16% of the votes. This indicates that it would be appropriate to make a number of different presentation styles available for the users to chose from.

A second comment by Scheuer *et al.* is of note, for although not directly applicable there is an equivalent sentiment for biology:

> our review has shown that the vast majority of the existing argumentation tools make use of graph-based argument representations. . . . But are graphs really the best way to visually represent arguments in educationally targeted systems? ([170] page 91)

This work examines only graphical styles of argument presentation. Furthermore the styles are based on pre-existing formats. The second quote from Scheuer *et al.* clearly demonstrates the limitation of this work. Overlooking this imperfection, this work shows that biologists find the Toulmin-like approach best; however, it is likely that a number of different styles should be provided in order to ensure that everyone is able to use a presentation format (s)he finds natural.

9.3 Which presentation mechanism - graphical or textual - seems most appropriate?

Although the prototype system presents arguments exclusively in a textual format, additional evaluation questions probe the differences in user attitudes to textual and visual presentation of arguments. In the first study the visual arguments are displayed only in the bottom-up style. As the second investigation shows this to be the least popular visual style, a comparison might be biased towards the textual arguments.

The comparison, see Section 8.1.3, seems to indicate that users are evenly split in terms of their preference. However, the visual arguments are perceived as being more readily understandable and marginally more intuitive - see Figures 8.6 to 8.9.

It seems that there is no single ideal solution, with the textual and visual arguments both required. Initially this appears surprising as the content of the graphical argument is identical to the textual argument - the difference is that the former is laid out in such a way that the individual elements and their relationship is clear. Perhaps the users, being new to argumentation, simply do not realise the significance of the diagrams and instead are interested purely in the text? It might be informative to repeat the study with users who are educated in basic argumentation skills - many US schools teach informal logic (or critical argumentation) and have done so for a considerable time whereas Scottish schools only recently have started to introduce a subset of these skills to the curriculum. The US biologists, more used to argumentation through their basic education, may provide a different result.

Alternatively, Verheij proposes the explanation may be concerned with the complexity of the domain and the information to be conveyed:

> A general question in the design of argument assistants is whether arguments should be graphically represented in the first place. Especially the complexities and subtleties of legal argument may impede such representations, and require natural language representations. . . . A compromise could be the dual representation of arguments, both graphically and in natural language. ([171] page 321)

This highlights the need for further investigation in the area. In the meantime it seems appropriate to include both textual and visual arguments.

9.4 How well do the standard argumentation concepts translate to the biological domain?

Biology has its own terminology, and concepts - these are refined and extended within the subdomain of gene expression. The same is true for most other domains, including argumentation. Therefore, can the respective concepts and jargon intermingle successfully?

The evaluation indicates that the biologists are able to cope with the argumentation concepts and terminology, with one notable exception. The confidence or believed strength of an argument, is indicated using the word “strength”. Unfortunately, in the gene expression subdomain the word is associated with the expression level - thus a percentage is taken to mean several different things.

A percentage could indicate the percentage of the tissue in which a gene is expressed. An alternative is the assumption that the percentage is referring to the amount of gene expressed in the tissue. In hindsight it is not difficult to see why the biologists misinterpreted the information. Fortunately an alternative term should not be difficult to obtain.

Ideally, such an issue would be identified during an early trial of the evaluation and thus rectified for the majority of users. Although trials of the evaluation were undertaken, the subjects were students at Heriot-Watt University, and thus not gene expression experts. Hence the trials did not detect the problem. As all four experts were tested on the same day, there was no opportunity to make a correction, thus all experts encountered the same limitation.

Clearly, this experience highlights the importance of precise terminology, and the value of conducting trial evaluations with a limited number of experts. Nevertheless, the lack of any major terminology or conceptual conflict demonstrates that argumentation works within this domain.

9.5 Can users interpret the information in the arguments correctly?

This query raises an interesting complication: what is the appropriate mechanism for judging the success of an argument? The simplest method is to say that a successful argument is one that persuades the audience. However, such a definition is not appropriate in this case as the goal is not to persuade a biologist that a gene is (not) expressed.

In the evaluation of Chapter 8, an attempt to understand the users' interpretation of the arguments is made - Section 8.1.3. However, it is debatable whether this could ever be accurate. The reason for the lack of clarity is the controversial nature of biology, as described by Jeffreys *et al.*:

> Different researchers interpret data in different ways, and even the same researcher may make inconsistent interpretations, adding an unreliable and non-uniform element to data processing. ([9] page 924)

As each biologist will interpret a piece of information differently, according to their own knowledge and views, it is difficult to know what a biologist should understand from a piece of information.

In Figure 8.8, from Section 8.1.3, the subjects are asked how intuitive they find a textual argument on a scale of 1 (not intuitive) to 9 (very intuitive). The median response is 5, suggesting that the users feel the textual argument is acceptable. However, examining the graph, it is clear that there is a wide spread of opinion ranging from 2 to 9. This implies that some users think they are able to understand and thus use the information, but that other users find it difficult to do so.

In conclusion, this question is very difficult to assess and it is unfair to provide an answer based on the limited evaluation carried out in this activity.

9.6 Do textual arguments provide the information necessary to reach a decision?

With respect to the discussion in Section 9.5, this consideration will disregard the complicating factors and use the results of Chapter 8 to explore a solution.

Figure 8.3 displays the answer to the question: *Was the amount of information presented in the arguments too much (1) or too little (9)?* The median response is 5, which is the perfect response indicating the arguments are providing enough, but not too much, information. Although Figure 8.3 is based on the responses from all users, it accurately reflects the thoughts of the 4 expert users who assigned values: 5, 5, 5, and 6. On the basis of this, the arguments appear to provide sufficient information.

9.7 Is the visual summary of the debate a worthwhile inclusion?

The original expert requested a visual summary of the argumentation process; however, he did not specify what this should be or the manner in which it should be presented. Section 7.6.2 describes the visualisation arrived at for this work. This section reports on the evaluation of it.

The final page of the system contains three distinct components. The first is a natural language summary of the result. The visual summary is the second, and the final element is the textual arguments. During the evaluation the users are asked which elements they apply when deciding if the gene is expressed in the tissue. Section 8.1.3 reports that fifteen of the eighteen users claim to employ the summary. This clearly indicates it is a useful element to include in any related or similar system.

Whilst the first analysis deals with the general notion of a visual summary, future discussions will relate to the actual presentation mechanism deployed. Section 8.1.3 demonstrates that the diagram confuses users. Only 44% are able to correctly identify which conclusion (expressed or not expressed) the diagram is indicating. Despite this users believe the diagram is easy to understand (Figure 8.10).

The users identify a difficulty in distinguishing between the different argument strengths. However, Table 8.8 shows that the majority of users are able to correctly gauge the strength of an argument by looking at the summary diagram.

In conclusion, the users claim to find the visual summary diagram easy to use yet do not interpret it correctly, and the users perceive a difficultly to exist in one area in contrast to the results of the evaluation. Overall, it seems that a summary tool is beneficial, but the current proposal requires further work.

9.8 A final comment

The open question at the end of the argumentation specific part of the first evaluation provided one very interesting comment:

> Would like to have more explanation of argumentation process and why one argument wins over another (see "Additional comments" from Section 8.1.3)

The most surprising aspect of this comment is that only one of the eighteen users made this point. Owing to the complexity of providing a good argument presentation, the implemented solution was less than ideal. As such, a significant amount of information was hidden from the end user including: the methods used to create and assign the degrees of confidence, the strategy applied to create arguments, and the approach used to settle conflict. Accordingly, similar comments were expected from a greater number of subjects.

9.9 Summary

This chapter used the evaluation contained in the previous chapter to discuss the best presentation of arguments and argumentation. The results from Chapter 8 were used to conclude:

- both textual and visual presentations of an argument should be included;
- a variety of different visual presentations should be provided so a user can chose the most appropriate;

- although difficult to produce, a visual summary of the argumentation is beneficial;
- the terminology and ideas from the argumentation community are transferable to the biological domain.

Chapter 10

Analysis: *Which form of argumentation?*

Following on from Chapter 9 in which the evaluation was used to explore the presentation of arguments and argumentation, this chapter considers the notion of *argumentation* in greater depth. It recalls Chapter 4, to consider the abstract question "*Which form of argumentation is appropriate as a solution to the above issues - and how effective is it?*" plus the associated subsequent questions:

1. What practical constraints exist? What are the consequences of these constraints? (see Section 10.1)

2. What is *argumentation* in this context? (see Section 10.2)

3. How can argumentation be implemented in this domain? (see Section 10.3)

4. Which architecture appears appropriate? (see Section 10.3)

Accordingly, this chapter shall examine which form of argumentation is appropriate in the domain of gene expression (Section 10.2), and thus the larger world of biology in general (Section 10.4). It shall examine the ways in which argumentation can be implemented (Section 10.3) and whether or not it can be successful. As part of this discussion a review of the constraints faced shall be conducted together with an account of the way in which they shape the solution developed in this work (Section 10.1). A summary will be provided in Section 10.5.

10.1 Constraints

Although there are many different forms and styles of argumentation, with an inclination towards computer science, this work takes a computational argumentation approach - that is it wants to generate arguments and debate automatically. However, this proves to be one of the main constraints placed upon this work.

Computational argumentation is still in the early stages of development. A set of guidelines that can be followed in order to create an argumentation-based solution is yet to be put in place. Furthermore Bench-Capon and Dunne [1] state that there is no single argumentation theory, and probably never will be:

> ...just as the attempts to construct a notional definitive non-monotonic logic from the disparate alternatives proposed in the 1980s are now recognised as ill-fated, such is likely to be the outcome of efforts to build an ultimate extension-based semantics. ([1] page 627)

Accordingly, work still focuses on developing theories rather than creating an optimised implementation. This explains why there is lack of available toolkits for performing argumentation. As discussed in Section 7.2.1, ASPIC was the only realistic choice for an argumentation toolkit when this work began.

At the time of writing the situation scarcely has altered. During the *ArguGrid* research project[1] *CaSAPI* [172] was wrapped in a JAVA extension to produce *MARGO*[2] [173], an argumentation engine focused on decision-making. Neither of these tools have updated their web pages since 2008, leading to the assumption that the development of both tools is now finished. Furthermore, MARGO clearly states on its web pages[3] that its target usergroup is “SCIENCE/RESEARCH”. A third argumentation toolkit is *ArgKit*[4]. This is produced by the developer of the ASPIC toolkit. As work is undertaken in the developer’s spare time progress is slow; however, it is still on-going. The idea is to provide a full implementation of Prakken’s framework, currently only the final layer has been completed. This effectively leaves the users to define all other layers: in the meantime it is unlikely that any biological project

[1] `www.argugrid.eu/`

[2] `http://margo.sourceforge.net/`

[3] `www.di.unipi.it/~morge/software/MARGO.html`

[4] `www.argkit.org/`

would go to the effort of doing this, preferring a ready-made solution. Other options exist, yet they seem to be academic tools designed for research not real world use. For example, Carneades is an argument mapping application comprising of two parts. The first is the GUI, the second an underlying argumentation engine originally written in the *Scheme*[5] programming language. Although Carneades may mature into something that is very useful, (at time of writing) its website states it is "currently under development". Noticeably, they are producing the latest version in *Clojure*[6], which renders it compatible with JAVA suggesting that this toolkit may hold promise for future endeavours.

Other constraints are associated with the world of bioinformatics. Firstly, there are a limited number of experts available to assist with projects like this. Furthermore, the time an expert can commit is restricted due to his/her many other commitments. Often the experts are distributed throughout many countries making it difficult to work with several experts. Fourthly, the online resources are often less integrated and harder to amalgamate than users expect. Even two obviously similar resources can have a raft of minor differences that cause problems. As an illustration, consider the two resources at the centre of this work.

Theoretically, EMAGE and GXD both use the EMAP anatomy. In reality, they each use their own slightly different version of the EMAP ontology: EMAGE has some tissues that are not in GXD and vice versa. Furthermore, these resources use different terms to mean the same thing - for instance, when describing a level of expression the GXD term *present* is equivalent to the EMAGE label *detected.* However, simply mapping terms does not necessary yield the correct result as demonstrated by the terms *absent* (GXD) and *not detected* (EMAGE). The EMAGE term means the gene is not expressed in the tissue or any of the component parts - child nodes in the ontology. In contrast, GXD's application of absent is such that the gene is not expressed in the tissue or in any component part, apart from the documented exceptions. In reality, the effort to map EMAGE to GXD is insignificant in comparison to mapping either resource to a third party such ABA - Chapter 12 explores this in more depth.

Although implicit in the previous text, it is perhaps worthwhile acknowledging

[5] `www.r6rs.org`

[6] `http://clojure.org/`

that formalising a large body of domain knowledge is challenging and thus expensive. Yet, often it is not essential to model the full extent of the knowledge, a notion captured by James Hendler's phrase "a little semantics goes a long way"[7].

In summary, a number of constraints shape the project and affect the solution. Creating a more effective argumentation-based proposition will require these issues to be tackled; however, it should be remarked that the constraints associated with the biological domain do not apply just to argumentation, but to all integrative technologies and proposals.

10.2 What is argumentation in this context?

As Chapter 5 explained, this work takes the view that argumentation is an automated debate in which arguments are generated, and compared. Argumentation is perceived as a device to generate and evaluate ideas, and thus encourage the users to consider more fully whether or not a gene is expressed. Argumentation fulfils both reasoning and explanation roles - new ideas are generated by inferring new arguments, and the system's reasoning is explained to the users by presenting the arguments to them. The implementation is tied, as remarked in Section 10.1, to the only available toolkit for performing argumentation.

One of the commonly perceived flaws raised by bioinformaticians, when this work is presented to them, is the lack of some form of probabilistic reasoning mechanism. There are approaches combining Bayesian nets and argumentation, e.g. Williams and Williamson use Bayesian networks to assess mammograms and then use argumentation to explain the system's conclusion to patients [81]. In William and Williamson's work, argumentation is merely a mechanism for constructing an explanation, it is not a reasoning process. Vreeswijk [174] tries to use Bayesian networks to perform argumentation. Mazzotta *et. al.* produce a similar system to Grasso *et al.* [11] which tries to persuade people to behave in a more health conscious manner [108, 175]. Mazzotta *et al.* use belief networks to model the persuasive power (strength) of each argument, this allows them to estimate which argument will be the most powerful and in turn tells them which argument to use. In essence, belief networks model the uncertainty

[7] `www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html`

of the strength of each argument.

Using a probabilistic mechanism would certainly help overcome a level of doubt in the minds of many bioinformaticians with regards to argumentation. The question remains: how to obtain the probabilities accurately? Seemingly, the solution is some form of machine learning. However, the appropriateness of this paradigm is questionable because of the limited availability of experts to help train the program, and the small number of available variables on which to construct the model. In addition, it is important to remark that the variables will change from resource to resource, thus there may need to be several networks, one dedicated to each resource. A detailed consideration of the surrounding issues is left as further work.

Regardless of how inferences are made, the application of natural language techniques to improve the clarity of the message seems an obvious extension. A body of expertise has been built up in the medical domain, where argumentation can be used to provide clear statements to patients or medics - see Section 2.6.2 for more details.

An alternative style of computational argumentation could be employed - possibly using the notion of a value-based argumentation framework (VAF) [78] in order to counter the issue of subjectivity. VAFs (see Section 2.4.5) use the concept of an audience to effectively create a new ordering, or preference, between arguments based on the opinions of each audience member. This ordering will effect the ability of arguments to defeat one another, and thus each audience member (biologist) may have their own specific result. The concern is that there does not seem to be any publicly available toolkit supporting the generation of a VAF.

It is plausible that the argumentation process could be re-envisaged as a dialogue between the computer and the end user. Grasso et al [11] try to persuade users to eat more healthily by using Perelman and Olbrechts-Tyteca's notion of schemes [17] to create suitable arguments whilst conversing with the user. This dynamic participatory vision of argumentation was never considered. Biologists (or bioinformaticians) run a tool (or tools), then analyse the output and finally reach a decision. Thus the approach of Grasso would be antithetical to the standard practice in the biological domain. It is impossible to state how a biologist would react to this radically different paradigm without a formal evaluation.

Although this work concentrates on computational argumentation, with a virtual debate being performed by the computer, there is room for alternative views of argumentation to be successfully applied.

10.3 How can argumentation be implemented in this domain?

Sharing the view of argumentation from the start of the preceding subsection, Chapters 5 to 7 recount the manner in which argumentation can be implemented in this domain. This particular discussion will focus on the perceived weaknesses of such an implementation.

The architecture in Chapter 7 aggregates data from two online resources, and uses the combined data as the basis for the argumentation process. However, there are a number of basic issues with this process.

Firstly, aggregating data requires the integration of the resources. As Section 10.1 remarks, often this is not straightforward, and can be impossible. Secondly, querying resources and pulling data may take a significant amount of time. If a substantial volume of data is obtained, conducting the argumentation may likewise be a lengthy activity.

The amount of time spent arguing depends on the rules and facts available. The volume of relevant data available in the resources cannot be controlled, accordingly the implementer must concentrate on the representation of that data. It is profitable to spend time constructing a well considered, and optimised, set of rules and corresponding portrayal of facts. However, such optimisation has its limits, and ultimately it may be necessary to increase the efficiency of the argumentation toolkit. This, in turn, may require an improvement in the underlying argumentation theory.

Decreasing the time spent querying the resources requires the initiation of some form of caching. As the resources regularly insert new annotations, and amend existing annotations, maintenance of this cache might be a sizeable task.

It is worth noting that any implementation will be tied to the argumentation toolkit utilised. Therefore, in order to truly answer the original question one must first establish which toolkit is currently most appropriate.

10.3.1 Which architecture seems appropriate?

In order to determine a suitable architecture, first it is necessary to decide upon the goal of the system. If the goal is to produce a proof of concept, or prototype, then the architecture presented in Section 7.3 is a logical starting point. It gathers all the required data when needed, and proceeds to argue, before presenting the results to the user. It represents an uncomplicated, reliable, mechanism by which to perform the task.

As discussed earlier in Section 10.3, the response time of such a system may be slow in cases where a considerable volume of data exists. Accordingly, the architecture implemented in this work would not be satisfactory in a production, or real world, environment. In such a situation, caching of biological data and possibly some common queries would be desirable. The addition of a cache would involve a minor change to the architecture, and have little effect on the workflow described in Section 7.3.1.

It can be difficult to integrate the separate resources in order to argue with the data they contain. Each resource will offer its own programmatic interface, which may or may not provide full access to the underlying data. If full programmatic access is not available then it may not be possible to pull the required data on the fly. Additionally, resources routinely upgrade and extend their programmatic interfaces, propagating change to the corresponding clients. In contrast, whilst the underlying repositories are continually having new data added, changes to their structure and/or organisation are far less frequent. With these considerations in mind, it becomes sensible to explore the potential of creating a local cache using a dump from each resource's repository. Many resources, e.g. MGI[8], publish a monthly dump of their entire repository (including GXD, gene functions, pathways, et cetera). Whilst that dump contains far more data than this work requires, the ability to create an optimised workflow for extracting, processing, and locally storing the necessary data is enticing. Furthermore, an increase in performance may be noticed as the access time for the resource is removed. Ultimately, the decision on whether to use a local cache will depend on the resources used and the access they provide to their data.

[8] `http://informatics.jax.org/`

With respect to the architecture, a cache would either replace or increment the existing databases. In terms of the workflow, whilst contacting the EMAGE database (or perhaps instead of doing so) the system could contact the local cache and proceed as per the rest of the workflow. If results are cached, rather than pulling data and arguing with it, the arguments could be retrieved from the cache and sent to the user.

One possible omission from the existing architecture is a web service providing programmatic access to cater for *in silico* experiments or perhaps the increasingly popular "app" platforms represented by the latest generation of mobile devices. The creation of such an interface to sit alongside the primary HTML interface presents no problems and shall not be discussed further.

An appropriate summary of this discussion should remark that although the architecture from Chapter 7 may require extension and refinement, it provides a suitable starting point for future work.

10.4 Is argumentation an appropriate mechanism for resolving the issues in this domain?

The final aspect, or subquestion, of question 1 relates to the value of using argumentation to tackle the topics of inconsistency and incompleteness described in Chapter 4. Can argumentation help a user unravel these issues and decide for themselves whether or not a gene is expressed? The solution developed during this work aggregates data from multiple resources and uses it as the foundation for an argument generation and evaluation process, the results of which are presented to the user.

There are clear benefits associated with the aggregation of the data. It reduces both the incompleteness and the number of resources a user has to visit. The one negative is that it increases the amount of inconsistency.

Argumentation can be used to infer likely results for not yet performed experiments, thereby reducing incompleteness further. For example, the following scheme allows gene expression information to be inferred for a stage when the adjacent stages have information:

There is no experiment E suggesting R for gene G in tissue T

Experiment $E2$ suggests result R for G in tissue $T - 1$

Experiment $E3$ suggests result R for G in tissue $T + 1$

$T - 1$ is the equivalent of T but in the previous Theiler Stage

$T + 1$ is the equivalent of T but in the next Theiler Stage

When the same result occurs in the stage before, and the stage after, it is extremely likely to be true for the stage in the middle

R is almost certainly true for G in T

In practice, for the current use case, there is only one implemented scheme that directly tackles incompleteness, and the corresponding rule is seldom applicable. However, there is a perceptible potential that may be exploited more effectively in another situation.

Upon the related problem of inconsistency, argumentation has a more dramatic effect in this use case. If desired, the argumentation may produce some form of decision as to which option (expressed or not expressed) is more likely according to its knowledge and the available facts. Additionally, there is value associated with the presentation of arguments without an opinion of expression. Simply exhibiting arguments allows the focus to be on the arguments, and thus the expert knowledge contained within. The reader can study the thought process of an expert, and understand what items of data the expert values when evaluating the gene expression data. Alternatively, if the reader is an expert, (s)he may access a wide range of relevant information, from numerous resources, far quicker than by manual means.

A convincing justification can be created for employing natural language based argumentation in most domains. For example, Green [12] uses the Toulmin scheme to generate text that provides powerful, clear, explanations for people undergoing genetic counselling. Green's work is applied to creating pamphlets; nevertheless, it could be adapted for use in an expert system.

This work concentrates on computational argumentation, as such the "argumentation" in the question refers to it, not the natural language argumentation of Green. The question asks if computational argumentation is appropriate - not if it is the "most" appropriate mechanism. However, both questions are considered.

Considering the explicit question, based on the experiences related in Chapters 7 and 8, the answer is positive. Computational argumentation can be successfully applied in a valuable manner: the ability to tackle incompleteness and inconsistency has been demonstrated. Moreover, the reasoning behind generated inferences can be explained to a biological user. Yet, to counterbalance this, the scalability of this approach is questionable.

The workflow documented in this thesis requires expert knowledge to be captured as schemes then turned into inference rules. These rules must be fed into an argumentation toolkit together with biological knowledge pulled from third party resources. Building a client to each resource can be expensive, especially if the resource routinely upgrades their programmatic interface. Documenting the expert knowledge is both expensive and difficult - Chapter 11 discusses this in detail. If a resource expands the type(s) of data it stores, the schemes/rules must be refreshed. The same is true if a new resource is added. Likewise, it makes sense to periodically review/refresh the schemes/rules to ensure they are still inline with current expert opinion.

Different approaches have not been compared in this thesis; therefore, it cannot answer with certainty the second question: is the current mechanism the most appropriate? However, informal feedback from bioinformaticians suggests that they would prefer a probabilistic solution. As discussed in Section 10.2, attempts at combining computational argumentation and statistical methods exist. The most obvious hybrid would use a statistical calculation to assess the strength of an argument, or its contributing domain information - Mazzotta *et. al.* [108, 175] do this within medical informatics.

Another related issue of interest, is whether or not this is a suitable area in which to apply computational argumentation. Although it is clear that other areas may provide more information, and thus more fuel for debate, there is no obvious reason why argumentation cannot be applied to the current domain. Yet, it must be remembered that the fruitfulness of that argumentation (and many other techniques) will be limited by the difficulty of integrating resources. To offset this, there is work proceeding into the integration of the aforesaid resources. For example, the creation of a coordinate-based reference space for the mapping and registration of neuroanatomical data [176] allows resources such as EMAP and ABA to map their individual anatomy models[9] into this new coordinate reference system, and thus each other. Finally, with the progression of such efforts resources may be truly integrated, ergo one limitation on the success of argumentation will be removed.

Whilst it must be acknowledged that the tools to support argumentation are currently too immature for real world use, and the principles behind those tools require refinement, both the metaphor of arguing and the practice of computational argumentation have potential uses within the life sciences.

10.5 Summary

During this chapter an analysis of the constraints affecting this work led into a review of the implementation from Chapter 7. Alternative methods for conducting computational argumentation are examined before a consideration of the appropriateness of other forms of argumentation. The culmination is a discussion of the applicability of computational argumentation to the current use case.

[9] ABA focuses exclusively on the brain, and so the entire anatomy model can be mapped. EMAP covers the whole anatomy, and thus only part of it can be mapped.

Chapter 11

Analysis: *Insights gained*

The third, and final, question from Chapter 4 is studied here: *What insights have been gained from the work, and how does that inform the future use of argumentation within biology?* This provides a single subquestion: *What foundations need to be laid before the common uptake of argumentation in biology can take place?* Additionally, Section 7.7 added the following:

1. How accurate are the schemes (and associated strengths) developed in Chapter 6? (see Section 11.1)
2. Is the expert used in this work truly representative of his community? (see Sections 11.1 and 11.2)

Chapters 9 and 10 contain a number of observations which have been exposed during this work, rather than restating them, this section concentrates on knowledge gained that does not fit directly into either of the previous sections. The goal is to provide a critique of the areas in which improvements can be made and potential pitfalls avoided.

This thesis has the objective of exploring argumentation within biological science. To do so it was necessary to adopt a workflow that supported the creation of a realistic system that might be applied within the chosen use case. The overarching ambition was never to produce a perfect solution, nor present a overly optimistic picture depicting argumentation as the solution to all the domain's imperfections. Instead the focus was on analysing the work undertaken, and striving to document the knowledge accrued in the process. During this chapter a thorough analysis of the workflow leading to the creation of the system in Chapter 7 shall be performed. This will highlight potential weaknesses, of the approach adopted in Chapters 5 to 7, and possible remedies. Most of the discussions in this chapter relate to the modelling of domain information: scheme creation (Section 11.1), the expert assignment of strengths (Section 11.2), and the conversion into logical rules (Section 11.3).

The fact that the end of this document is able to provide such an analysis demonstrates that the earlier work was successful - the use of argumentation within biology has been investigated. Whilst this chapter makes clear that computational argumentation needs to evolve, there is an undoubted potential to be exploited by future work. Section 11.4 discusses the future of argumentation in biology before the summary in Section 11.5.

11.1 Scheme creation

Ultimately, the schemes can only be as good as the mechanism that creates them. Subsequent examination of this process reveals a number of issues from which future work can benefit. As the workflow centres on the biological expert, so too do the observations in this section.

The original expert's job was to review experiments to produce annotations, which are stored in the EMAGE database - the actual results are hidden away in journal articles. As such, he performed the task the schemes model on a daily basis. Furthermore, as the senior editor, the expert had to supervise the other editors and where necessary resolve differences of opinion. His expertise for the chosen use case was unquestionable; however, the expert had no previous experience of participating in a knowledge capture process, and found it difficult to engage with. As this discussion makes clear, the extraction of expert knowledge can be very challenging.

There is little published research on the creation of argumentation schemes by a domain expert with which to compare the current work. The available literature on schemes often focuses on the dialogue and natural language aspects.

Silva *et. al.* [177] is interesting because it starts with already documented "reasoning templates" (effectively diagrammatic schemes) and asks a domain expert to use those to explain his reasoning, customising them if necessary, during case-based reasoning. The act of customising the schemes provides a mechanism to help the expert describe his knowledge. No such pre-existing schemes were available for the current domain.

Shipman and Marshall [178] discuss the difficulties associated with working with a number of knowledge formalisms, including argumentation and the Toulmin scheme. Although their work does not focus directly on the application of schemes a number of their ideas do transfer over, for example:

> Tacit knowledge is knowledge users employ without being conscious of its use [179]. Tacit knowledge poses a particularly challenging problem for adding formal structure and content to any system since, by its very nature, people do not explicitly acknowledge tacit knowledge ([178] page 342).

The notion that experts are not aware of all their own knowledge presents a massive impediment for all knowledge-based approaches. Similar ideas can be found in the work of Bliss [180], which suggests that experts develop *mental models* of concepts and processes that can be very hard for them to access. Such knowledge, referred to as *deep knowledge*, is unarticulated. Knowledge that has never been articulated can be extremely difficult for the expert to recall [181, 180, 182]. The biological expert was being asked to provide tacit knowledge, and struggled.

Bliss [180] suggests a further hurdle: mental models naturally evolve as the person gains experience and knowledge, or as a result of the person consciously thinking about their activities and knowledge. Consequently, the act of asking an expert to document their knowledge can change that knowledge. This could explain why the expert continued to modify the schemes. Essentially, the schemes only capture a snapshot of the expert's ever-changing mental model. The captured knowledge will expire eventually, and thus the schemes need to be refreshed regularly.

In addition, Shipman and Marshall quote Mittal and Dim [183]:

> We believe that experts cannot reliably give an account of their expertise: We have to exercise their expertise on real problems to extract and model their knowledge ([178] page 34)

Essentially, if experts are trying to think about how they use their expertise, they are not thinking about the expertise. In effect the task becomes thinking about how to provide the knowledge required rather than actually providing it.

Mittal and Dims' quote demonstrates that real world examples should have been used from the outset, instead of introducing them in meeting 4.

The task facing the expert was made harder for him by the inexperience of the analyst, and the analyst's comparative lack of biological knowledge. This caused the expert to simplify the discussions to a relatively high level - as commented on in Section 6.3.2. Clearly it cannot be stated for certain that an evaluator with greater biological knowledge would have generated more biological schemes.

On occasions when an attempt was made to capture biological ideas, the challenge was often too arduous. An example of this was encountered in the third meeting, and was the reason behind the change of approach requested by the expert. During the second meeting, the expert had suggested that the area of the spatial mappings could provide a measure of their quantity. Schemes 3, 4, and 5 from Appendix A.2.1 record the idea that the greater the number of individual voxels[1] in which the gene is expressed, the more likely the gene is to be expressed. Furthermore, the higher the *morphological match*[2], and *data pattern clarity*[3], the more probable the voxel mapping is to be correct. In the third meeting, the expert was asked which of the three measures was most important - the goal being to create an ordering which could be used as the basis of the degree of belief. This ostensibly simple question caused the expert to ask his colleagues for assistance. They decided that a fourth measure, *model quality*[4], is most important, but concluded that further research was required to rank the remaining three measures. Additionally, the editors believed that a new investigation was required to identify a level at which a certain number (or percentage) of voxels being expressed is meaningful.

[1] The 3D models are comprised of Volumetric Pixels (voxels).

[2] Indication of how well the EMAGE 3D model correlates to the subject of the experiment. It is a score between 0 and 3, with 0 being terrible and 3 very good.

[3] Indicates how clear the experimental result is, again this is measured with a score between 0 and 3.

[4] The lower the quality of EMAGE's model, the larger the voxels and thus the less meaningful each voxel is.

These schemes were removed, and a less elaborate replacement sought. The expert was able to state that a spatial annotation can be dismissed when the percentage mapped is less than 3% (scheme 7 from Appendix A.2.1). Yet, when asked to quantify his assertion that the meaning of each percentage is weakened as the size of the tissue increases, he felt that further experimentation was required. The scheme now says that a spatial annotation is more meaningful than a second spatial annotation, if it affects a larger area (the 6th scheme in Appendix A.2.1). Although this second scheme appears to be a simplification of a complex matter, it is accurate, and as precise as was possible.

The fundamental restraint, for both of the previous two examples, is a lack of biological knowledge. There are gaps in the community's knowledge, hence the expert's knowledge is not complete.

A further topic of interest, is that several of the expert's comments are open to interpretation. For example, when the expert suggested that textual annotations are more reliable than spatial annotations, was that a rule that should be implemented as a scheme? Or was it a remark that related to the ordering of the schemes for those annotations, and thus implied the relative degrees of belief for the annotations? Although it is believed that both approaches are valid, the former was selected. In effect the statement was treated as a rule in which the expert has a degree of confidence. This may have been because the analyst was concerned with creating schemes, and so had a propensity for interpreting the remark as such.

A final analyst related concern is the difficulty of managing a library of schemes. Almost 70 schemes exist, and keeping track of these schemes and their interactions is a sizeable task, as demonstrated by the existence of a small number of holes in the knowledge, i.e. missing schemes. For example: what should happen when the same resource produces conflicting conclusions for the same experiment[5]? There is a related issue: when should the knowledge recording stop? Biology is a vast, interrelated series of subdomains, therefore it is easy to cross boundaries and take tangential paths. Moreover, it is not difficult to get lost in detailed discussions. It is not clear where the

[5] This can happen in a very rare situation in which a gene is said to be not expressed in a high level tissue, but is described as being present in a substructure. Due to propagation, if the gene is expressed in a child node, it is also expressed in the parent.

boundary between enough and too much knowledge lies. For example, although one scheme relies on the idea of two annotations being from the same experiment, there is no scheme that indicates how to identify this. Is this an implementation issue that can be ignored, or an important piece of biological praxis that should be recorded?

The last expert related quandary identified is possible bias. In particular, scheme 5 from Appendix A.2 suggested the annotators of EMAGE are more reliable than the annotators of GXD. This may be a correct statement; however, with no independent expert to verify it, there is no way of ensuring its accuracy. Although other experts are available, they too work for the EMAGE project - so the bias may not be removed through their inclusion.

The potential bias raises an interesting question. Should the expert ever be overruled? If the goal is to capture his knowledge, surely the answer is no? However, if there is a clear bias should it not be “watered down”? The latter approach is employed in relation to scheme 5 from Appendix A.2, with the subsequent (6^{th}) scheme suggesting that a disagreement between the EMAGE and GXD editors should be a reason to distrust both annotations. Yet this is in stark contrast to the previous claim that the schemes would be a model of the expert’s reasoning. The justification of this decision is not transparent. However, it is felt that including such a scheme without asking for the opinions of the GXD curators is unacceptable. Furthermore, the expert approved the weakened scheme, meaning it is logically no different to the other schemes.

One possible solution to the issue of bias is to employ multiple experts. Additionally, including multiple experts may reduce the impact of evolving mental models by excluding the most controversial (i.e. likely to change) ideas. Lindgren [184] is working to create a decision support system for medics trying to diagnose dementia. To aid this, together with her colleagues, Lindgren is building a framework that allows medics to work collaboratively to create schemes [185]. Crucially, the medics are able to work from fully researched and published medical guidelines, which means bias has already been resolved. Obstacles such as tacit knowledge have likewise been removed, as the medics merely convert the guidelines from one form (published document) into another (list of schemes). Without the existence of biological guidelines, the current task is substantially different to that facing Lindgren’s medics.

Although the basic idea of collaborative scheme generation ostensibly seems promising, it raises some intriguing questions. Would allowing the expert the opportunity of brain-storming ideas with another expert furnish more biological schemes? Would it have helped the expert to dispense knowledge that he was unable, or unwilling to do? Is it fair to believe that it would identify areas of personal opinion as opposed to generally accepted opinion and scientific fact? Furthermore, what difficulties would it in turn cause?

Shipman and Marshall contribute to this discussion:

> The difficulties of creating useful formalizations to support individuals are compounded when different people must share the formalization. . . . Differences occur not just within a group of users but between groups as well. ([178] page 344)

This raises the prospect of agreement between those that work as database curators; yet, conflict between that group and those who carry out the experiments. Might this even go further, with disagreement between subgroups? For example between industrial researchers and academics, or industrial scale researchers and traditional small scale research labs. Clearly, some form of managed conflict resolution would be beneficial. The fact that the experts may be in different geographical locations suggests that the entire process of scheme creation may need to be based online. The above concerns require further scrutiny.

11.1.1 Comments from the evaluation

In addition to the preceeding analysis, the evaluation in Chapter 8 advances reasons to doubt some of the schemes. The evaluation is designed to test some of the schemes the expert provided in Chapter 6. In particular, the expert proposed that the users' opinion of the research team behind an experiment, and the journal that published it, would impact on the users' confidence in the experiment. For this to be true, the users must recognise the research team and the journal. Clearly, the non-biologists were unable to help.

Table 8.5 summarises the responses of the subjects upon being asked to supply their confidence in journals related to the current use case. The standard deviation for each journal is relatively low, this suggests that the users have a similar opinion of that journal. Furthermore, the average (mean) scores for the journals do not vary widely, indicating that the users' believe the journals to be of broadly similar quality. These points seem to suggest that a scheme based on the users' confidence in a journal is redundant.

There were 9 biologists, including 4 domain experts. Yet Table 8.6 demonstrates that most research teams (authors) are recognised by just 2 users. Again, this raises qualms over the relevance of the associated scheme, it is likely that questions relating to confidence in the research team should be hidden by default, with an option allowing domain experts to answer.

The intuitiveness of the arguments (schemes) is called into doubt too (Figure 8.9). This time their author rated them as 8/9 - suggesting they were very intuitive. However his colleague, and fellow EMAGE editor, responded with 3/9. As these individuals perform almost the same role within EMAGE, it seems natural to suppose their reasoning would be aligned far more closely. This contradiction leads towards the conclusion that different biologists use different methods and patterns of reasoning, which supports the claim that using only one expert for the creation of schemes is insufficient.

11.2 Expert assigned strengths

Expert assigned strengths are measures of the confidence (degree of belief) the biologist creator had in each scheme. The aforesaid biologist disagreed with those strengths during the evaluation, which raises concerns over the reliability of their capture.

As for the scheme's strengths, they are mainly unremarkable. However, as Section 6.3.2 comments, the expert employs a small number of the values available to him. This connotes that the mechanism of associating a percentage with the scheme is too finely grained, prompting the question: what else could be utilised?

Silva *et. al.* [177] use symbolic methods to quantify the strength of an argument. Essentially, their expert created a range of labels (e.g. "nuisance", and "problem") which are assigned numerical values (nuisance = 6.0 and problem = 7.0). Then symbols (so-called *qualitative signs*) are used to vary the final numerical score. For example, ++nuisance = 6.5. Such an analysis is based on the work of Fox and Parsons [186] in the medical domain, indeed the symbols used $(++, +, -, --)$ are very similar to a notation $(+++, ++, +)$ routinely used as a form of shorthand on patients' medical records. Fox and Parsons [186] compare a range of numerical and symbolic dictionaries for use in their own argumentation implementation, logic of argumentation [74]. Earlier work by some of the same authors, [187], awards a score of ± 1 or ± 2 to each argument scheme, which is inherited by the generated arguments, depending on whether the argument is strong (2), weak (1), for (+) or against ($-$) the conclusion being argued over.

In this document, the strength of an argument is used in two ways. Firstly, the argumentation engine uses it to resolve conflict between two arguments. Secondly, it provides the end users with some indication of how reliable an individual argument is. For the latter to be true, the mechanism for communicating the strengths must be meaningful to the end users. The use of common medical shorthand makes sense if creating a system for the medical fraternity. However, biologists do not use such codes. Therefore no symbolic language is immediately available.

The use of natural language labels seems far more intuitive and user friendly. Yet two issues exist here. First, is the expert's definition of a term the same as the users'? The resources, and the experts who work on them, use different labels as there is no standardised vocabulary. Therefore, a level of vagueness and unreliability will be introduced. The second issue is implementation related.

The ASPIC argumentation engine uses a degree of belief to capture an argument's strength - this is a floating point value between 0 and 1. Regardless of what mechanism is used to capture the strength, it ultimately needs to be converted into a numeric value. Such a mapping will need to be performed by the expert. Consequently, the expert will have to assign numerical values to the schemes irrespective of whether he initially assigns symbols or labels. Adding labels, or symbols, seems to add complexity and work, for this reason, the expert was asked to deal entirely with numbers. The simplest numerical abstraction appeared to be the idea of percentages, and so this was employed. Additionally it has the advantage of being very easy to convert into the floating point number that the argumentation engine stipulates.

11.2.1 Expert reliability

An issue to be aware of when dealing with experts, is the accuracy of the individual specialists. In this work an expert is used both to provide the schemes, and to assign a degree of confidence to them. It is difficult to know how to proceed without such input. Yet, as Walton notes, the users of expert opinion are often not capable of judging the quality of that opinion and thus simply apply it:

> It is quite common for presumptions to be based on expert opinion where the person who acts on the presumption - not being an expert - is not in a position to verify the proposition by basing it on hard evidence within the field of expertise in question. ([55] page 39)

Walton goes on to caution that such experts, or authorities, may not prove to be correct:

> But even when they are nonfallacious, as used in a dialogue, appeals to authority are generally weak, tentative, presumptive, subjective, and testimony-based arguments. They are inherently subject to critical questioning, or even rebuttal, on various grounds - especially on grounds relating to the reliability of the source cited. ([55] page 34)

Walton's opinion is backed by a number of scientific studies, as Hansson reports:

> Experimental studies indicate that there are only a few types of predictions that experts perform in a well-calibrated manner. Thus, professional weather forecasters and horse-race bookmakers make well-calibrated probability estimates in their respective fields of expertise [188, 189]. In contrast, most other types of prediction that have been studied are subject to substantial overconfidence. Physicians assign too high probability values to the correctness of their own diagnoses [190]. Geotechnical engineers were overconfident in their estimates of the strength of a clay foundation [191]. ([192] page 35)

These quotes combine to illustrate the difficulty in using expert opinion - often it is not correct, and the user has no way of knowing what (s)he can trust. Again, this seems to advocate the need for a collaborative approach to scheme generation in which multiple expert opinions balance one another, and thus remove prejudices and errors.

Jeffreys *et al.* highlight that an extra range of complications must be considered when dealing with biologists:

> Different researchers interpret data in different ways, and even the same researcher may make inconsistent interpretations, adding an unreliable and non-uniform element to data processing. Once the data has been interpreted, the reasoning behind an interpretation may be lost or only vaguely recalled. When a researcher leaves a research group, the method used to interpret data goes with them and is lost. Finally, the researchers interpretation may be biased towards getting a preconceived result. ([9] page 924)

The quote from Jeffreys *et al.* accentuates the importance of subjectivity to biological decision making: "Different researchers interpret data in different ways". The initial expert's views on strength differ with those of his colleagues, and the other experts in the evaluation. Such subjectivity affects not just the strengths, but the schemes too.

11.3 Implementing schemes as rules

Making allowances for possible weaknesses within the scheme creation process, another area of difficulty is the translation of schemes to rules for use in the argumentation toolkit.

The expert who helped in this work was an individual with no knowledge of logic and no desire to learn it. Consequently, although the expert created and then verified the schemes he did not, and could, not verify the resulting logical inference rules. Accordingly, there is the possibility that the rules do not match the intention of the expert.

Although constrained natural language, the schemes are still natural language with all the lack of precision that entails. Hence, a number of practical implementation decisions had to be taken when making the rules. For example, one scheme states that when EMAGE and GXD have different textual annotations (one says the gene is expressed, the other not expressed) for the same experiment that is a reason to doubt both annotations. Although clear in meaning, this scheme does not state how to decide when two annotations are derived from the same experiment. Nor does any other scheme, which highlights a lack of depth to the schemes produced. In this instance, the decision was made by comparing the citation information for experiments.

Despite Verheij [62] publishing a seemingly clear mechanism by which to perform the necessary transformation from scheme to logical rule there is still an element of doubt because there is an aspect of subjective interpretation. Consider the following scheme:

Spatial annotation SA suggests verdict V
Spatial annotations are a reasonable indication of expression
Therefore it is likely that V is true

1. Is SA reliable (correct)?
2. What do the textual annotations show?
3. How good is the morphological match?
4. How high is the pattern clarity?

It has four critical questions which are manifestly straightforward. Consider question 1. The content is clear: if the spatial annotation is not reliable then this scheme cannot be applied. Yet the implementation is not. Should this be an exception that stops this scheme being applied when the annotation is not valid? Could it be that a condition must be added forcing the spatial annotation to be valid? Might this point to another argumentation scheme (or schemes) that consider(s) the validity of spatial annotations? Although further guidance was sought in the literature, it was not found. Perhaps this is because the solution is obvious to those with a (informal) logic background that dominate the domain.

In particular, the difference between extra conditions and exceptions seems minor, surely these are the proverbial "two sides of the same coin." The former favours prevention, and the latter cure; however, the end is the same: an argument is stopped.

Examining the distinction from the implementation angle presents a different perspective. The extra conditions stop the argument being generated; however, the exception allows the argument to be created and then kills it. Taking the argumentation engine's behaviour into account - in allowing the argument to be created and then killed, the argument will appear as a defeated argument within the output. Yet, if an argument is never constructed, due to extra conditions, it will not. Instead the only sign of what happened will be buried deep inside the optional log, and thus lost. Accordingly, modelling as an exception seems to be more explicit. Importantly, a balance must be struck between providing users with all the necessary information, and overburdening them with too many arguments.

While it is undemanding to implement the critical questions once they have been categorised, performing the classification is more problematic, and often seems to rely on personal opinion.

In summary, there are three distinct areas of concern associated with this translation process:

- The schemes are written in natural language and thus are open to interpretation;
- The mechanism for converting schemes to rules is not as explicit as initially thought;
- The expert is not able to verify the logical inference rules.

One possible solution to all of these issues is the use of a framework designed to help experts create rules for use in an argumentation system. Lindgren is working towards creating a system to help physicians diagnose dementia [193]. The reasoning undertaken by the system [184] is an evolution of the argumentation system created by Parsons *et al.* [186], using Walton's notion of schemes and critical questions. Lindgrens' work has the problem of which method to use to capture expert knowledge for use in a rule based expert system. The chosen solution is to develop a collaborative framework to allow medics to document clinical guidelines as schemes, and convert the schemes to rules [185].

There are clear parallels between Lindgren's problem and the situation documented here. However, as Section 11.1 reports, there are disparities between the current use case and Lindgren's. Lindgren is working in the medical domain, and thus her experts are interpreting a set of published rules (clinical guidelines) and documenting them firstly as schemes and secondly as logical rules. In gene expression no such guidelines exist. Consequently, the biological expert employed here had to generate his own guidelines and views, which is a formidable task. Moreover, Lindgren's experts seem to have been working together at the same computer - it is not clear whether her approach would work if the experts are geographically dispersed. Despite these obstacles, if ultimately proven successful, Lindgren's solution may provide a template for a similar approach in biology.

11.4 Foundations for future argumentation work in biology

Throughout this thesis an exploration of argumentation in biology has been conducted. The work has taken one form of argumentation theory and applied it to one biological use case. By evaluating the approach taken, this thesis has demonstrated that argumentation can be successfully utilised in the life sciences. Additionally, analysis of the evaluation has provided a range of future extensions and/or alternative practises that could be used to improve upon the approach documented here. This section will gather the various threads, from Chapters 9, 10 and 11, to create a record of what needs to be improved/changed in order for computational argumentation to flourish. There are two separate aspects to this current task. Firstly, what needs to happen within the biological domain, and secondly the challenges facing the argumentation community.

Starting with the biological domain, it is clear that integration of resources is essential. Furthermore, the case study emphasised the importance of providing standards for improving the capture and documentation of provenance information. For bioinformaticians to implement argumentation based solutions, they will require the argumentation community to develop reliable standardised methods and toolkits.

Related to the above comment regarding argumentation methods and toolkits is the need for the community to make these items more readily available. Unfortunately, *argument* is a very common word, so simply searching online for information is often too general. For example, entering *argumentation biology* as a search term into Google[6] on the 31st August 2010 resulted in "about 11,600,000 results". The first result was the definition of *argument* from an online biological dictionary[7]. The second was a series of articles presenting arguments for/against stem cell research[8]; likewise, the third result was an online discussion of evolution[9]. The fourth was a link to a paper [10] discussing natural language scientific arguments in science education. Although the fourth result was argumentation, in the wider sense, it was not computational argumentation. Changing the search term to *computational argumentation biology* presented "about 754,000 results", the majority of which dealt with computational biology, not argumentation.

Other tasks for the argumentation world include tackling the open question of how best to present arguments and argumentation, and the development of a collaborative mechanism to enable biologists to develop and model the knowledge themselves. These last tasks are interdisciplinary requiring input from social technologists to ensure the environment is natural for the biological end users.

Noticeably, this chapter has concentrated on issues related to modelling the domain information. It appears that the effectiveness of computational argumentation within biology hinges on the quality of the domain modelling. Regardless of the application realm, the effort required to model domain knowledge is significant. This prospective cost presents a substantial barrier to the successful adoption of computational argumentation. Yet the same is true for the Semantic Web, and as James Hendler and others [194] have stated - a little semantics goes a long way.

[6] www.google.co.uk

[7] www.biology-online.org/bodict/index.php?title=Argument

[8] www.helium.com/knowledge/1388-arguments-for-and-against-stem-cell-research

[9] www.biology-online.org/biology-forum/about15556.html

11.5 Summary

Chapter 11 concentrates on the laboriousness of domain modelling for the current use case. It explores the knowledge capture and transformation processes, highlighting the difficulties and discussing possible remedies. This chapter concludes by stating that for computational argumentation to prosper within biology it is necessary to better integrate biological resources and improve upon the number, and quality, of argumentation tools and supporting methods available.

Chapter 12

Argudas: an evolution

Following on from the positive evaluation, discussed in Chapter 8, a grant proposal was submitted to the BBSRC to create a tool based on the research contained within this thesis. This was duly accepted, and work began on *Argudas* in spring 2010. Dr. Albert Burger was the grant holder; however, the work was primarily driven by the author, with assistance from Dr. Gus Ferguson who coded the user interface, and helped with the evaluations.

Although this chapter is the penultimate chapter of this thesis, in actuality the work described here was conducted in parallel with the maturation of Chapters 10 and 11. Appropriately, there is a relationship between these three chapters in which developments, and changes in understanding, related to one chapter impacted upon the others. The current chapter is provided last because it facilitates a better presentation of the narrative.

Argudas is designed to be an evolution of the work described in earlier chapters. For that reason the system's mechanics will not be discussed further. Instead this chapter centres on the issues that have affected the work. The chapter starts by reviewing the ways in which Argudas improves on the previous work (Section 12.1), subsequently it examines additional problems identified during the development of Argudas (Section 12.2), before reviewing the final system (Section 12.3). Concluding remarks are made in Section 12.4.

12.1 Improving on previous work

One issue clearly highlighted by the evaluation (Chapter 8) was the notion of subjectivity. A second matter raised during informal discussions with typical users was the breadth of information provided. These are two areas in which Argudas attempts to improve on its predecessors. Priority was given to the former point, as it was likely to have the most impact.

12.1.1 Reducing subjectivity

In order to make Argudas seem less subjective to the end users two main elements were amended. Firstly, the summary of the result was removed from the results page. Secondly, the schemes and their associated degrees of confidence were reviewed. Each alteration shall now be explored and justified.

12.1.2 Reducing subjectivity: removing the conclusion

Pre-Argudas prototypes presented a one line summary of the debate at the top of the results page, e.g. Figure 7.12 states "the arguments appear to suggest the gene is expressed". The evaluation, Section 8.1.3, demonstrates that the expert users were divided with regards to this summary. Half of the experts used it, the other half did not. A plausible reason for this is provided in the previously quoted text from Jeffreys *et al.* [9]: "Different researchers interpret data in different ways".

The phenomenon of subjectivity is explored, in relation to argumentation, in the philosophical writings of Perelman and Olbrechts-Tyteca [17]. Perelman and Olbrechts-Tyteca introduce the notion of an *audience* to capture the idea that each member of an audience has their own reasoning process, and thus each member of the audience will judge the same argument differently. This means that there is little point in the system trying to decide whether or not the gene is expressed. Instead the system must generate arguments for and against the gene being expressed, and allow the users to evaluate these arguments in order to reach their own decision. In effect, the system should aggregate and evaluate data, presenting the relevant data to inform the users' decision making process.

Although the summary is a one line epitome of what the argumentation shows, users often interpret it as a decision. Accordingly, they expect it to match their own reasoning. However, as Jeffreys *et. al.* and Perelman and Olbrechts-Tyteca show, this is not realistic because users apply their own knowledge and beliefs, and therefore may reach a different conclusion. In such an event, the summary provides a reason to distrust the system, and effectively acts as a barrier between the system and the users. That being so, Argudas does not present a one line summary. Instead it aggregates, and interprets, information before recapping it.

12.1.3 Reducing subjectivity: schemes and degrees of confidence

As remarked in Sections 11.1 and 11.2, a number of evaluation subjects (including the original expert) disagreed with the schemes and the degrees of confidence assigned to those schemes. Argudas did not have the resources to create a new set of schemes as this was a substantial task; nevertheless, it was possible to review the degrees of confidence. To this end, the original expert and one of his colleagues were asked to review the entire list of previously generated schemes and award them a score:

0 disagree with the scheme;

? don't know - scheme is very weak and is on the border between being rejected and being classified as a weak scheme;

1 weak scheme, i.e. low confidence;

2 moderate scheme, i.e. medium confidence;

3 good scheme, i.e. high confidence.

In total, the two experts were asked to assign a score to sixty-eight schemes. The experts completely agreed upon - that is they gave exactly the same score to - sixteen schemes. A further thirty-three schemes were assigned a similar score. The notion of *similar* being defined as an adjacent score, i.e. if one expert assigned **2**, then either a **1** or a **3** would be classified as similar. If the two experts assigned scores that were neither adjacent nor exact matches, they were deemed to disagree - this happened with 19 schemes.

In conclusion, the experts broadly agreed on 72% of the schemes. This left 28% of the schemes for which the disagreement was substantial. Regrettably, the original expert (the EMAGE senior editor) emigrated shortly after this exercise was completed and was no longer available to assist in the development of Argudas. Therefore this disagreement was never resolved, nor was its root cause investigated.

Potentially the source of the disagreement was very interesting, as it was not clear whether the conflict between the experts was caused by a genuine difference of opinion or a difference of interpretation. As the schemes were written in natural language, the latter is a distinct possibility.

12.1.4 Extending Argudas for richer argumentation

Argudas aims to improve on previous work with the integration of further resources - more resources means extra information and the prospect of richer arguments. Initially the microarray data contained in ArrayExpress was targeted. Unfortunately, this highlighted a number of integration issues.

Firstly, ArrayExpress does not use the EMAP anatomy ontology. Secondly, accessing the data held by ArrayExpress was difficult as they did not provide direct programmatic access to their database. At that time access was via a RESTFUL web service, which provided limited functionality and did not allow access to the data required for this work. For example, initially it was impossible to ask for all the genes expressed in a healthy mouse's pancreas at stage 24 because ArrayExpress did not compute multi-factor statistics. That is, they computed which genes were expressed in the pancreas and which genes were expressed in stage 24 separately and there was no way of presenting the intersection at that time. The team behind the resource was working on improving this interface and claimed that such functionality would become available in the future; however, the delay was problematic because of the time constraints associated with Argudas. Finally, ArrayExpress had less data for the developmental mouse than expected: only three stages were covered. Weighing the costs and benefits it was decided not to pursue this integration further.

As work on ArrayExpress stopped an investigation of the Allen Brain Atlas (ABA) and GENSAT began. Both of these resources are databases of *in situ* experiments focusing predominately on the adult mouse's nervous system, i.e. brain, spinal cord, *et cetera.* The latter project makes available a full database dump. The former supplies an extensive range of RESTFUL[1] interfaces that provide access to the desired information.

However, bringing the data from these two new resources into Argudas is problematic. Neither resource uses the EMAP anatomy - as both focus on the brain they have a finer granularity for the brain tissues than EMAP. Hence it is necessary to attempt some form of mapping from their respective anatomies to EMAP. Moreover, these resources use their own measures to describe the level of expression, GENSAT natural language terms and ABA floating point numbers, which must be mapped across to the corresponding EMAGE/GXD terminology.

Mapping between the different anatomy ontologies employed by the resources is based on a series of alignments produced by Jiménez-Lozano *et al.* [195]. As both GENSAT and ABA have a finer granularity than EMAP, mapping from those resources to EMAGE/GXD results in a loss of precision.

The second task is straightforward for GENSAT as their choice of labels is similar to EMAGE's[2]. Whereas EMAGE has *not detected*, *detected*, *weak*, *moderate*, and *strong* GENSAT has *undetectable*, *weak signal*, and *moderate to strong signal.*

Mapping EMAGE/GXD expression levels to their ABA equivalents is more complex. There are three separate measures of expression level published by ABA. Firstly there is the raw experimental information, secondly there is the average information (across all the experiments for a particular gene and tissue), and finally there is a mathematical aggregation of the expression level and expression density. For current purposes, the first class of information is most suitable. The ABA numbers must be mapped to the EMAGE/GXD natural language descriptions. This is achieved by using a list of mappings previously generated by ABA. These "mappings" are a series of cut-offs that determine whether the expression level is *not expressed*, *weak*, *moderate* or *strong.* There are different limits for each part of the brain. In order for the limits to be applied to the tissues lower down in the anatomy hierarchy, the limits need to be propagated through the brain in a similar manner to the gene expression information.

[1] `www.ibm.com/developerworks/webservices/library/ws-restful/`

[2] Which, in turn, is very similar to GXD's.

Once this work has been undertaken it is necessary to determine what level of integration is appropriate for these resources. At the most basic level it would be possible to merely report the results contained in ABA and GENSAT. If either of these resources agreed with an annotation from EMAGE/GXD, it would increase confidence in that annotation. Fully integrating ABA and GENSAT would require the generation of schemes for these resources, which is substantially more work and would necessitate involvement by a resource expert. In the case of ABA such an approach may not be fruitful; ABA does not publish all the data it collects, accordingly many of an expert's schemes may not be usable. The restricted resources of Argudas meant that only the former option was realistic.

Although an interested biologist may raise a number of concerns regarding the anatomy and expression level mappings described above, currently there is no better way of aggregating data between the four resources of interest.

12.2 Changing the notion of arguing

As Argudas was developed it became clear that the number of arguments generated varied enormously. For some queries there are no annotations and therefore no arguments. With other queries over ten annotations were retrieved from EMAGE and GXD, accordingly a large number of arguments were generated. For example, arguing for *bmp4* - future brain in stage 15 generated two hundred and fifteen arguments. Clearly, no biologist would read all the arguments, hence there could be no guarantee that (s)he would read all the important information. This realisation led to the conclusion that the potential number of arguments was too high, and steps were taken to reduce it.

12.2.1 Reducing information overload

When Argudas generated a large number of arguments they were all unique in terms of their content (wording, and order of words), yet semantically several arguments seemed to duplicate one another. Identifying semantically equivalent arguments is not a minor task. The definition of *equivalent* seems to depend on the individual using the system and the biological task they wish to perform.

There are a number of common interpretations and actions that are not appropriate for certain biological tasks, and which individual biologists may, in general, reject. For example, the EMAP anatomy ontology is defined using part-of relationships. The outcome of which is that positive levels of expression are routinely propagated up the ontology to higher level tissues; for example, if *bmp4* is weakly expressed in the telencephalon, it is normally correct to say that *bmp4* is weakly expressed in the future brain. Nevertheless many, but not all, biologists prefer direct annotations over propagated ones, thus if a second annotation suggested *bmp4* was not detected in the future brain, the second annotation would take precedence.

Likewise, there is a similar problem with the granularity of information desired. Finding two distinct annotations with the same conclusion is a powerful argument for trusting the conclusion. However, the granularity of information desired affects the decision as to whether or not the annotations are in agreement. Assume there are two annotations: one annotation suggests *bmp4* is strongly expressed in the future brain, and a second annotation demonstrates *bmp4* is weakly expressed in the future brain. If the biologist is attempting to determine if the gene is expressed or not expressed, then these annotations may be taken to agree. Yet, if the aim is to determine the level of expression, these annotations are conflicting.

The goal of reducing the number of possible arguments was further hindered by a user request for more positive aspects to be highlighted. For instance, although an argument was created when the experiment's probe information was absent, no argument was created when it was present.

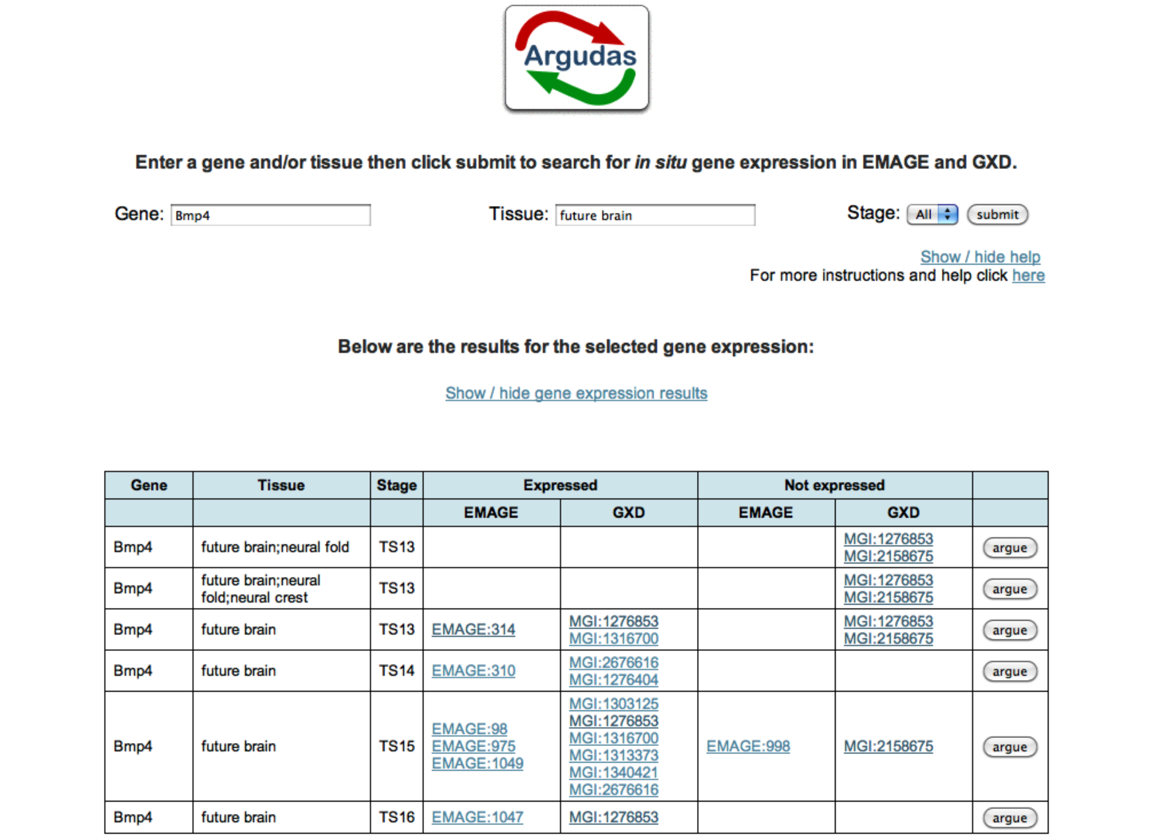
In summary, Argudas' users appeared to wish for a broader range of prospective arguments, and yet a smaller number of realised arguments. Reconciling these competing aims seemed improbable, until someone remarked that the problem was not the volume of the arguments but the amount of text to be read. It transpired that a significant number of potential users wanted to scan information rather than read it.

12.2.2 The notion of argument reconsidered

Previous work, and the initial version of Argudas, used the ASPIC argumentation engine to generate and evaluate arguments inside a virtual debate. These arguments were presented to the users as a natural language paragraph - this display mechanism was chosen as it was the preference of the original expert. However, feedback suggested this choice was subjective (Section 8.1.3). Furthermore, the potential for a large number of arguments to be generated appeared to imply that the initial approach was sub-optimal. There was a clear need to find an alternative method for displaying arguments.

During internal project discussions it was proposed that the argumentation mechanism should be reconsidered. This approach was based on the belief that users wanted quick access to certain key attributes of the annotation. Theoretically there was no need to employ the argumentation engine to create and evaluate arguments. Instead, the most important schemes (as identified by the scheme revision process), should be the basis for a range of key attributes that describe the annotation. The schemes indicate whether or not the information stored in EMAGE/GXD should increase or decrease a users' confidence in an annotation. As such, Argudas should extract information from the resources and present it, with associated key highlights, to the users. It then becomes the duty of the users to evaluate the information.

In order to test this hypothesis two mock interfaces were created and evaluated. There were three steps to each interface. The first two steps were the same: select a gene and/or tissue of interest; report on the available annotations and allow the users to ask for more information if desired. Figure 12.1 shows both of these: initially the query is *bmp4* - future brain in all stages; the query causes all combinations of the gene and tissue to be displayed in a table. The table presents all relevant annotations, summarises what each annotation shows, and provides a link to the resource's web page for that annotation.



Enter a gene and/or tissue then click submit to search for *in situ* gene expression in EMAGE and GXD.

Gene: Bmp4 Tissue: future brain Stage: All submit

Show / hide help
For more instructions and help click here

Below are the results for the selected gene expression:

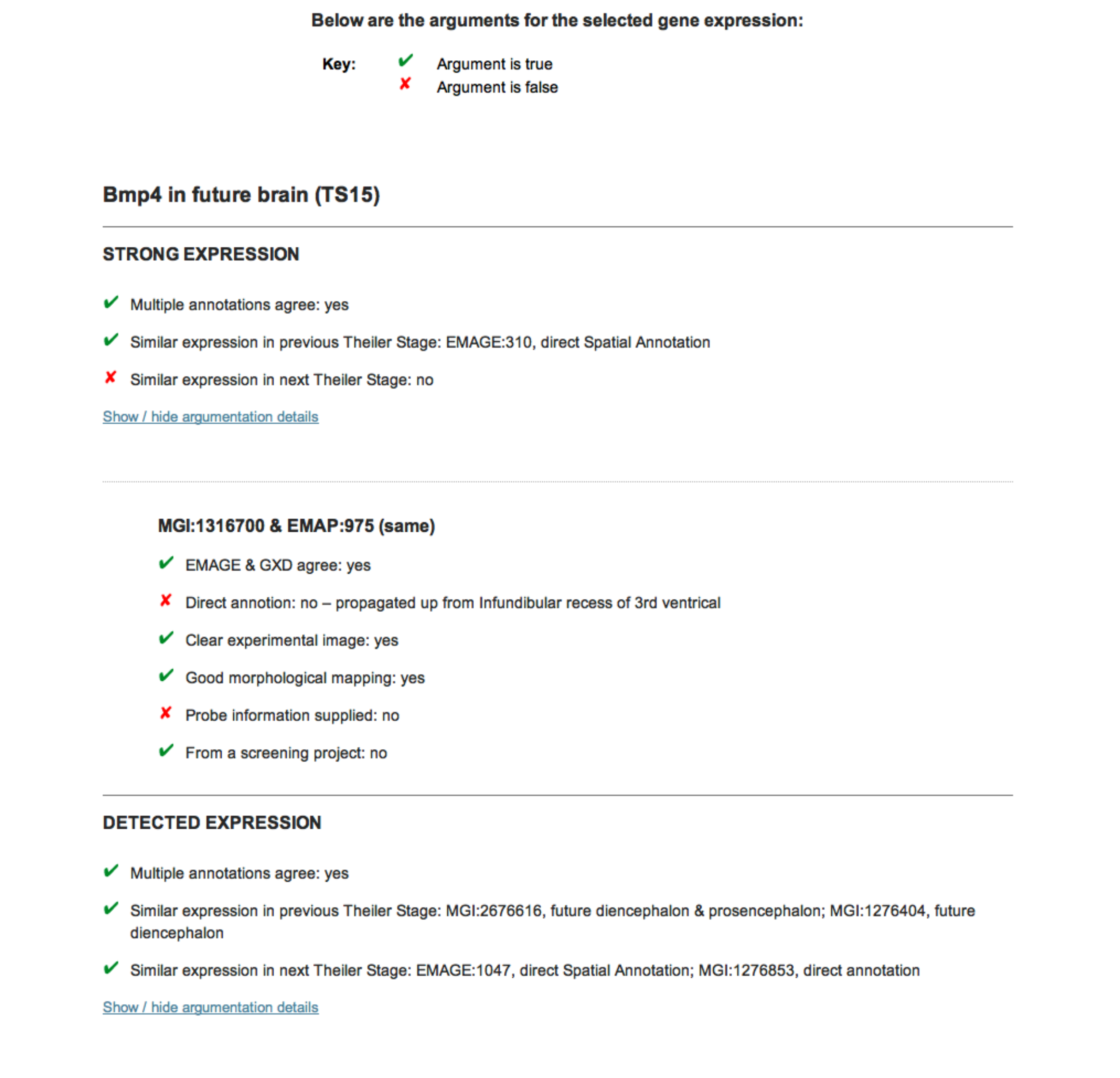
Show / hide gene expression results

Gene	Tissue	Stage	Expressed		Not expressed		
			EMAGE	GXD	EMAGE	GXD	
Bmp4	future brain;neural fold	TS13				MGI:1276853 MGI:2158675	argue
Bmp4	future brain;neural fold;neural crest	TS13				MGI:1276853 MGI:2158675	argue
Bmp4	future brain	TS13	EMAGE:314	MGI:1276853 MGI:1316700		MGI:1276853 MGI:2158675	argue
Bmp4	future brain	TS14	EMAGE:310	MGI:2676616 MGI:1276404			argue
Bmp4	future brain	TS15	EMAGE:98 EMAGE:975 EMAGE:1049	MGI:1303125 MGI:1276853 MGI:1316700 MGI:1313373 MGI:1340421 MGI:2676616	EMAGE:998	MGI:2158675	argue
Bmp4	future brain	TS16	EMAGE:1047	MGI:1276853			argue

Figure 12.1: Mock-up of the user interface: simple form allows user to search for gene and/or tissue in relation to a particular Theiler Stage. Doing so produces a table summarising the relevant annotations found in EMAGE and GXD.

In some situations the table in Figure 12.1 would be enough to resolve a biologist's question; i.e. it is clear that *bmp4* is expressed in the future brain in stage 14. On the occasions when the table is not helpful, or does not provide enough information, clicking the *argue* button provides more information.

In the first mock interface a number of textual arguments were displayed - in a similar manner to Figure 7.12. The second interface can be seen in Figure 12.2 - the arguments are now a list of key attributes such as *multiple annotations agree*. Whether or not an attribute should strengthen a users' confidence in the annotation is indicated with a tick or cross. The attributes are divided into two layers - first by expression level, and then by annotation. For each level of expression there are three attributes that indicate how likely that level of expression is. Asking for more information causes the second layer of attributes to appear. This allows the users to evaluate the annotations individually, and collectively as a group that promotes a specific expression level.



Below are the arguments for the selected gene expression:

Key: ✓ Argument is true
✗ Argument is false

Bmp4 in future brain (TS15)

STRONG EXPRESSION

✓ Multiple annotations agree: yes

✓ Similar expression in previous Theiler Stage: EMAGE:310, direct Spatial Annotation

✗ Similar expression in next Theiler Stage: no

Show / hide argumentation details

MGI:1316700 & EMAP:975 (same)

✓ EMAGE & GXD agree: yes

✗ Direct annotion: no – propagated up from Infundibular recess of 3rd ventrical

✓ Clear experimental image: yes

✓ Good morphological mapping: yes

✗ Probe information supplied: no

✓ From a screening project: no

DETECTED EXPRESSION

✓ Multiple annotations agree: yes

✓ Similar expression in previous Theiler Stage: MGI:2676616, future diencephalon & prosencephalon; MGI:1276404, future diencephalon

✓ Similar expression in next Theiler Stage: EMAGE:1047, direct Spatial Annotation; MGI:1276853, direct annotation

Show / hide argumentation details

Figure 12.2: Mock-up of the user interface: potential display mechanism for arguments. Clicking on an *argue* button from Figure 12.1 produces an output where different attributes are assigned a positive (tick) or negative (cross) indicator. Positive indicators demonstrate the annotation is more likely to be correct.

Informal exploration

The mock interfaces were explored with the assistance of two expert users from the *Medical Research Council*'s *Human Genetics Unit* (HGU). One expert had participated in the evaluation (Chapter 8) and in the revision of argumentation schemes and confidence values discussed in Section 12.1.3. The second expert had attended a presentation on Argudas, but previously had not been involved in any work associated with this thesis.

Each expert user was presented with a description of the planned evaluation, then a structured walkthrough was conducted. Using a protocol, the user was guided through each interface using the same search example: *bmp4* - future brain - stage 15. They were asked to raise any issues or aspects they liked or disliked while undertaking the interface evaluations. The experts were then asked to score the interfaces out of ten in terms of their usability. Ultimately, a limited set of questions was asked to determine the users' opinions of the requirements for refining aspects of the interface and argument presentations. In particular, one question directly asked which style of presentation the subject preferred. A second question asked how their favoured approach could be improved.

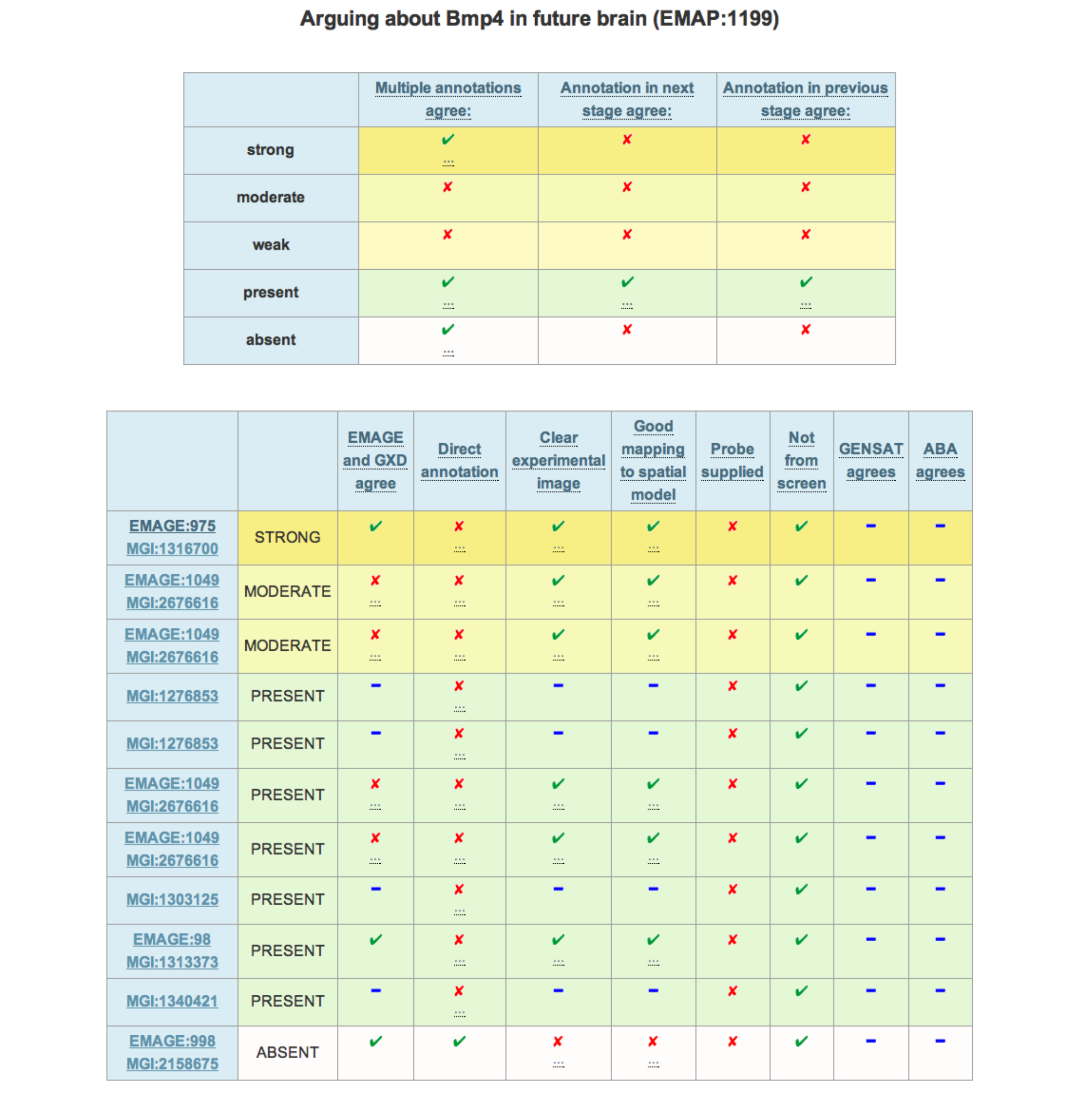
Although the exploration was too limited to allow any statistical analysis, the qualitative data collected provided useful indications for future development of the system. During this small informal exercise, the test users were kept apart; however, they reached identical conclusions:

1. the revised interface, using key attributes rather than textual arguments, was better;

2. a further improvement could be made by placing the content of Figure 12.2 into a table.

The experts differed in their implementation of the tabulation. One believed the existing expression level layer of attributes was acceptable, and that only the annotation layer should be converted into a table with one table for each expression layer. The other expert user preferred all the expression level layer attributes in one table, and all the annotation layer attributes in a second table.

Despite being limited, the evaluation demonstrated that the second interface style, in which arguments become attributes, was preferred over the previous version.



Arguing about Bmp4 in future brain (EMAP:1199)

	Multiple annotations agree:	Annotation in next stage agree:	Annotation in previous stage agree:
strong	✓ …	✗	✗
moderate	✗	✗	✗
weak	✗	✗	✗
present	✓ …	✓ …	✓ …
absent	✓ …	✗	✗

		EMAGE and GXD agree	Direct annotation	Clear experimental image	Good mapping to spatial model	Probe supplied	Not from screen	GENSAT agrees	ABA agrees
EMAGE:975 MGI:1316700	STRONG	✓	✗ …	✓ …	✓ …	✗	✓	–	–
EMAGE:1049 MGI:2676616	MODERATE	✗ …	✗ …	✓ …	✓ …	✗	✓	–	–
EMAGE:1049 MGI:2676616	MODERATE	✗ …	✗ …	✓ …	✓ …	✗	✓	–	–
MGI:1276853	PRESENT	–	✗ …	–	–	✗	✓	–	–
MGI:1276853	PRESENT	–	✗ …	–	–	✗	✓	–	–
EMAGE:1049 MGI:2676616	PRESENT	✗ …	✗ …	✓ …	✓ …	✗	✓	–	–
EMAGE:1049 MGI:2676616	PRESENT	✗ …	✗ …	✓ …	✓ …	✗	✓	–	–
MGI:1303125	PRESENT	–	✗ …	–	–	✗	✓	–	–
EMAGE:98 MGI:1313373	PRESENT	✓	✗ …	✓ …	✓ …	✗	✓	–	–
MGI:1340421	PRESENT	–	✗ …	–	–	✗	✓	–	–
EMAGE:998 MGI:2158675	ABSENT	✓	✓	✗ …	✗ …	✗	✓	–	–

Figure 12.3: Final presentation mechanism for arguments. Clicking on an *argue* button from Figure 12.1 produces an output where different attributes are assigned a positive (tick) or negative (cross) indicator. Positive indicators demonstrate the annotation is more likely to be correct. The first table covers expression level attributes, with the second table providing details on the individual annotations.

12.3 Argudas: the final system

Following the investigation discussed in Section 12.2.2, in-house experimentation was conducted by the author in conjunction with Dr. Ferguson. After due consideration, the second expert's approach was adopted - see Figure 12.3.

Figure 12.3 represents the final presentation mechanism. The important schemes have become columns. Strengths of expression are featured as rows in the first table. A tick indicates a positive reason to trust the level of expression, a cross provides a reason to doubt the level. The three dots underneath a tick or cross indicate that more information is available as a mouse-over, e.g. the tick for *strong - multiple annotations agree* presents the experiments with matching annotations. The second table has one row dedicated to each annotation. The strengths of expression are colour co-ordinated with the first table. The key is the same, with the addition of blue dashes reporting that data is unavailable.

The top table in Figure 12.3 is fixed in size - one row for each level of expression. The bottom table in Figure 12.3 varies in length, according to the volume of information available. The greater the number of relevant annotations, the more rows the table has. Whilst there is the potential for the bottom table to become very large it does not happen currently. Due to the incompleteness of the domain, in the vast majority of cases there are considerably fewer annotations (less rows) presented than shown in Figure 12.3. Of course, the underlying resources continue to grow, and in the future the size of the bottom table may become an issue. However, it is worth remarking that biologists are used to dealing with long outputs, for example BLAST, and are adept at quickly scanning for the desired information.

The conclusion, from Section 12.2.2, that key attributes should be presented in tabular format is very significant. Implementing the recommended tabulation has a major side effect: there is no need to perform computational argumentation. As such, there is no need for the argumentation engine. Instead the users now perform the argumentation themselves using the aggregated and curated information presented by Argudas.

One consequence of this is that only those with a knowledge of the use case are able to use Argudas, which makes it more restrictive than earlier work. Offsetting this, is the positive way in which the change helps address the issue of subjectivity. This occurs because the users are now able to apply their own confidence values and criteria to the decision making process. Furthermore, because the users are able to build their own arguments, Argudas becomes a more flexible tool. For example, it is now possible to use Argudas to decide where, and at what level, a gene is expressed.

12.3.1 Changes to the architecture and workflow

The removal of the ASPIC argumentation engine, and the change in the presentation mechanism are only two of the improvements to previous work. The inclusion of data from ABA and GENSAT requires a significant amount of pre-processing in order to conduct the alignments. Furthermore, GENSAT has no real time programmatic access point. Caching data locally is the only option.

With a local cache already a necessity, it seemed prudent to explore the potential of caching data from EMAGE and GXD. Whilst the population and maintenance of this cache creates a significant workload, it greatly improves access times to Argudas. All the experimental data, from all four resources, is now cached.

The removal of the argumentation engine significantly simplifies both the architecture and the sequence diagram from Chapter 7. Essentially, Argudas is now a client-server system that presents information contained in a database. When a user enters a query, Argudas contacts the local cache and gathers all relevant experimental data. Data is analysed according to the knowledge contained within the argumentation schemes, aggregated and finally presented.

12.3.2 Evaluation

Evaluation of the second, and final, version of Argudas involved testing the user experience with a number of biologists, bioinformaticians, and computational biologists working with/for the HGU. The functionality of Argudas had already been evaluated with previous versions of the user interface, so this evaluation was aimed solely at evaluating the new version of the user interface and the method of presenting results.

Method

In total eight individuals took part in the exercise over the course of a single day. The evaluation was performed on site, at the HGU, using a Chrome browser running on an Apple Macintosh laptop computer.

Due to the busy schedule of the evaluation participants, the evaluation was kept to twenty minutes. The first half was a structured walkthrough, in which the users were guided through the process of a typical query. The second half consisted of a standardised usability evaluation questionnaire and a number of more open-ended questions to gather the users' opinions on the proposed presentation of results. The protocol can be seen in full in Appendix G.

Results

The evaluation of Argudas was too small to draw any statistical conclusions. However, the evaluation questionnaire presents a good summary of the tool's usability - see Figures 12.4 to 12.8.

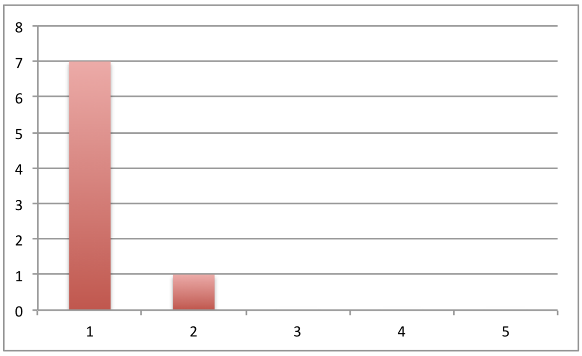


Figure 12.4: Please rate the response time, excellent (1) - bad (5)? Median = 1.13.

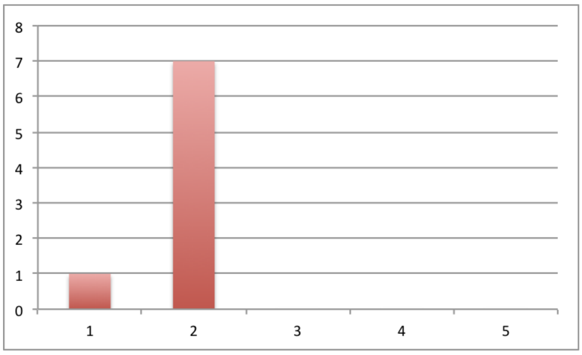


Figure 12.5: Please rate the appearance, excellent (1) - bad (5)? Median = 2.

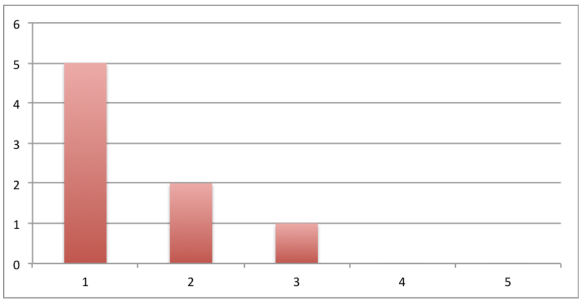


Figure 12.6: How easy did you find the first table to understand, very easy (1) - very difficult (5)? Median = 1.

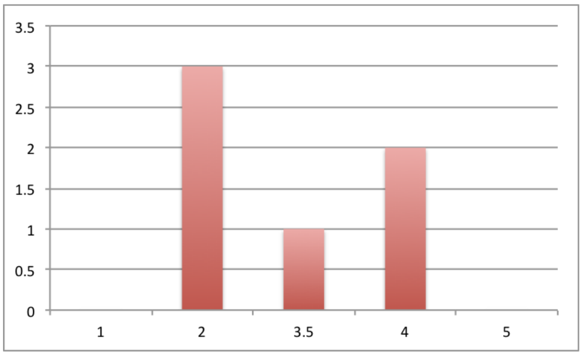


Figure 12.7: How easy did you find the second table to understand, very easy (1) - very difficult (5)? Median = 3.

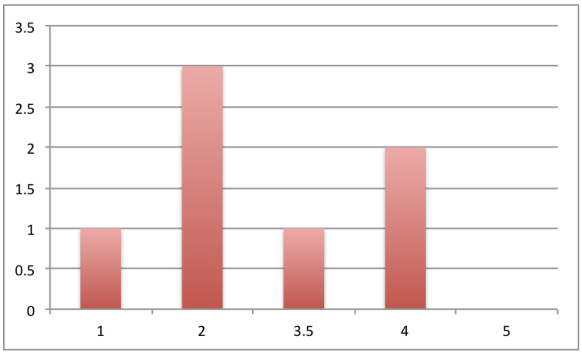
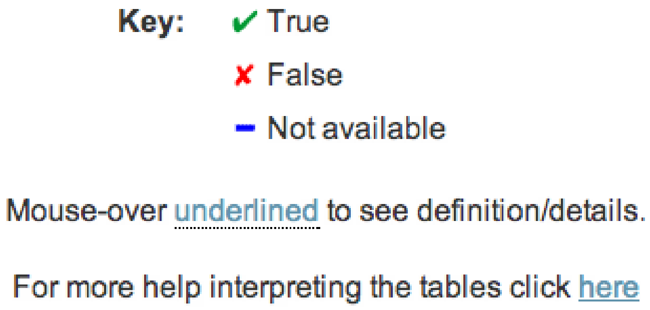


Figure 12.8: How easy did you find the third table to understand, very easy (1) - very difficult (5)? Median = 2.5.

Discussion

Argudas presents its output inside three tables. The first is presented in Figure 12.1. The second is at the top of Figure 12.3, and the third is at the bottom of Figure 12.3.

In contrast to the first Argudas table, the second and third tables received relatively low scores in the evaluation (compare Figure 12.6 to Figures 12.7 and 12.8). It is believed that this performance is related to two minor design issues highlighted by the evaluation. Firstly, almost all users managed to miss the key (see Figure 12.9) that was displayed above the second Argudas table, and thus were confused by the blue dash (the dash, or hyphen, can be seen in the second table in Figure 12.3 where it appears inside cells that do not have a cross or tick). Secondly, few users were able to discover the mouse-overs without assistance. The second issue is related to the first, because the mouse-overs were described within the key. However, it is also caused by the non-intrusive design of the mouse-overs. The three dots, used to indicate the presence of a mouse-over, were deliberately subtle because of a desire not to overpower the associated tick/cross. While this aspiration was undoubtably met, it transpired that many users simply did not see the dots. Clearly, there is a balance to be struck, and the current approach requires enhancement.



Key: ✓ True
✗ False
– Not available

Mouse-over underlined to see definition/details.

For more help interpreting the tables click here

Figure 12.9: The key that explains how to interpret the tables in Figure 12.3.

Further experimentation will be required to resolve the mouse-over matter. Making the key more prominent is a less challenging task. This can be achieved by simply moving it from above, to beside, the tables it describes. Alternatively, having the key float alongside the tables, as the user scrolls through the page, is an option.

Whilst additional work is needed, there is no doubt that the latest version is an improvement on earlier interfaces. By allowing users to “eyeball” information, Argudas provides a quick way of surveying and assessing all available information.

12.4 Summary

Argudas - a real world tool based on the research contained within this thesis - is the theme of this chapter. The discussion recounts the evolution of Argudas from a system using computational argumentation towards a system that utilises notions, such as argumentation schemes, from argumentation theory. As part of this transformation, the presentation of information shifts from a traditional argument-centric approach, to a tabular format bespoke to Argudas. The revamp is in direct response to user feedback and is a reflection of the users’ changing attitudes to the system.

Ultimately, the system generated reflects the wish of the users to quickly scan the range of available information. Accordingly, Argudas is a success from the users' perspective.

Argudas demonstrates that the metaphors of *argument* and *argumentation* are very powerful; however, the associated technology has not matured sufficiently to merit adoption. In essence, Argudas is an "argument" for the *future* application of computational argumentation in real world biological applications.

In summary, this chapter echoes the message of the Chapter 10: the tools and methods supporting computational argumentation need further development, yet the metaphor of arguing can be used within the life sciences already.

Chapter 13

Conclusion

Simplicity and naturalness are two characteristics of argumentation, which render it comprehensible and thus make it a powerful tool. This is the insight with which this work begins. Subsequently, argumentation is applied within a representative biological use case in order to determine the merit of applying argumentation to the wider life science world.

The final chapter of this thesis reviews the narrative of the journey, from conception to evaluation, documented within this thesis. Pivotal stages along the path are recalled, before ultimately summarising the contributions of this work.

Section 13.1 starts the chapter by outlining the use case. Next, Section 13.2 provides a précis of the content of this document before the key contributions are reviewed in Section 13.3. Section 13.4 contains a brief exploration of potential future work, before Section 13.5 provides a conclusion to this work.

13.1 Recap of use case

Whilst it seems apparent that argumentation can be applied within biology, it is appropriate to test such a belief. This thesis does just that, focusing on a use case from the sub-domain of *in situ* gene expression for the developmental mouse.

The developmental mouse is the mouse from the moment it is conceived, until the moment before it is born. The changes that occur over this period are documented in a series of so-called *developmental stages*. These stages are used to assign a time to the gene expression information captured through experimentation.

DNA is the body's blue-print. It tells the body what to build, how to build it, and when to do so. DNA is split into modules called *genes*. Each gene is responsible for a small deed, accordingly genes need to collaborate in order to achieve significant outcomes. Via a complex series of chemical reactions the instructions contained within each gene are activated and de-activated at different times during the development of the mouse.

An *in situ* gene expression experiment aims to determine to which areas of the mouse's body a particular gene is contributing. Furthermore, the activity of the gene is tied to a time through the use of the developmental stages. Due to the complex nature of the domain, results cannot be guaranteed to be accurate. Nor can the interpretation of the results (so-called *annotations*) be trusted automatically. Accordingly, contradictions are common. Moreover, not every gene has a complete expression profile because not every experiment has been performed.

Experimental results are spread across a number of resources, which are individually, and collectively, incomplete and inconsistent. This situation is repeated across a number of sub-domains in the life sciences, making the chosen use case representative (in this particular aspect) of the life sciences in general.

13.2 Thesis synopsis

This thesis starts by describing argumentation and the subset of it devoted to computational argumentation. It paints a rich picture of a domain in which the notion of a reason to believe something (i.e. an *argument*) can be used as a unit of reasoning within a debate (i.e. *argumentation*). These basic notions may be applied to tasks which are seemingly unrelated, or they can be utilised to model human-like interactions and reasoning processes.

In situ gene expression for the developmental mouse is the use case for this work. The critical characteristics - from the perspective of this work - are the inconsistent, incomplete and distributed nature of the underlying information.

When considering the best mechanism for tackling the issues associated with the use case, three elements of argumentation immediately appeal. Firstly, the ability to debate the accuracy of a particular conclusion: for example, the gene is (not) expressed. Secondly, the use of arguments, created during the debate, as explanations of the reasoning process. Finally, the idea of automating the above concepts in order to create a system that can reason over the underlying information and help biological users resolve the inherent limitations of the use case.

With this overview, it becomes necessary to consider the exact form of an argument: what is it a reason to believe? Contemplating the domain's information: biological researchers perform an experiment, documenting the result as an image. This is interpreted to produce a series of annotations: e.g. the gene is weakly expressed in the brain. Collectively this information may be published in a journal. Additionally, or alternatively, it will be published by one or more online resources. These resources, for example EMAGE and GXD, attempt to verify then publish the annotations. Furthermore, they publish the image that documents the result along with some basic provenance information.

Biologists use resources like GXD to discover where a gene is expressed. This means the overall goal of the argumentation should be to do likewise. An expert biologist will examine the individual annotations, attempting to gauge their validity, and ultimately arriving at a conclusion as to whether or not a particular gene is expressed in a specific tissue. The argumentation process mimics the real world. As a result, arguments are reasons to believe that a gene is, or is not, expressed in a particular location. To further define the notion of argumentation for this situation, it is necessary to consider the relationships between the different types of information available.

Experiments are the basis for the journal articles, which are often the foundation of the entries in EMAGE and GXD. However, the journal articles are unstructured data sources, and although many are now open access, without text mining their content is out of reach. Arguments must originate solely from the content of resources like EMAGE.

Delving deeper, the annotation is essentially supported (or justified) by the other information (image and provenance) held in the database. Conflict may occur between two experimental results (i.e. images) or between the interpretations of those images. As there is no automatic way to analyse the images, the former is ignored. Argumentation may be used to debate which interpretation is most likely to be accurate. Additionally, argumentation may be used to postulate an annotation when the necessary experiment has not been undertaken.

In order to automate this debate, a third party argumentation engine is applied. It consumes knowledge of how to interpret the domain information, together with domain information, before delivering arguments. EMAGE and GXD supply the domain information, thus only the knowledge of the best means to analyse that information is missing. As this knowledge is confined to a limited number of experts, at least one of this number must convey their understanding of how to construe the biological information.

The expert employed in this thesis was the senior editor of EMAGE. His knowledge is captured and documented using Walton's notion of schemes. An *argumentation scheme* is essentially a natural language rule with premises and a conclusion. When the premises are true, the conclusion is likely to be valid too. In order to determine the appropriateness of applying an individual scheme in a given context, they are commonly associated with a number of critical questions. When asking a question, a negative answer makes the application of the scheme doubtful.

Following a number of meetings, and iterations, the expert's knowledge is documented within schemes. Additionally, each scheme is assigned a so-called *degree of belief*. This is a score which indicates the expert's confidence in the scheme. The higher the value, the higher the confidence. Eventually, the schemes are used to generate arguments, to which the scores are passed on. These scores allow the argumentation engine to settle disputes - when two arguments are contradictory, the argument with the higher score wins.

The argumentation engine indirectly applies the schemes. Before schemes can be used, they have to be converted into the engine's own language - a PROLOG-like logic. This is achieved by following the method Verheij published. His technique involves converting the natural language inference rule into the desired logic, and then using the critical questions to add extra conditions, contradicting rules or exceptions.

With a basic outline established, a couple of prototypes were implemented in order to evaluate the approach and the underlying notion of argumentation. The basic workflow, created in this thesis, entails extracting information from the domain sources before feeding that information into the argumentation engine beside the converted expert knowledge. In response to a query, the argumentation engine generates arguments, which are presented to the user.

While the above workflow, and related architecture, remain largely static during this work, the presentation of arguments and argumentation evolves. It is necessary to adapt the presentation for the typical end user; however, this is a challenge.

All of the systems built in the lifetime of this document were evaluated. Some formally, others informally. On occasions the evaluations focus solely on the presentation of arguments and argumentation, at other times the value of the concepts of arguments and argumentation are explored. Additionally, the reliability of the expert knowledge, and his assignment of scores, are dealt with.

The evidence from these evaluations is combined with the experience gained during this work to produce an analysis of the suitability of computational argumentation within the life sciences. This analysis is formed in parallel to the creation of Argudas.

Discussion of Argudas - a real world tool attempting to tackle the above issues in the current use case - concludes the thesis. It builds upon the understanding obtained during this work to produce a platform that allows a quick inspection and consideration of the information held within EMAGE and GXD. Due to the timing constraints and the difficulty of presenting arguments, computational argumentation is currently not the most appropriate technology for the system biologists envisage. Accordingly, it is removed from Argudas.

Despite the frustrating conclusion to Argudas, the evidence from this work suggests that argumentation will be eminently useful within the biological domain once its full potential is realised. The simplicity and accessibility of argumentation render it a powerful metaphor that will undoubtably increase in value and popularity as the complexity of computational endeavours in the life sciences intensify.

13.3 Review of contributions

During this thesis the following contributions to academic knowledge and the life science community are made:

- Creation of argumentation schemes to document expert knowledge relating to the analysis of gene expression information;

- Development of an architecture for an argumentation-based platform, which aggregates distributed data before arguing with it;
- Investigation of the mechanisms with which argumentation can be presented to a biological audience. Encompassing evaluation, analysis and direct comparison of a number of presentation forms including natural language representations and graph-based visualisations;
- Exploration of a gene expression use case leading to a consideration of the application of argumentation in the biological domain including an assessment of the role of the wider notion of argumentation theory plus the application of computational argumentation;
- Review of the issues that hinder the development of argument-based solutions within biology;
- Proposal of ideas for the development of both argumentation and biology to enable further penetration of argumentation-related technology into the biological domain;
- Implementation of a real world tool that implements, evaluates, and evolves the ideas discussed in this work;
- Frank consideration of the future role of the notion of argumentation within the life sciences, including the genesis of a number of key questions regarding the future of this technology.

In general, there is a lack of experience papers describing the use of argumentation. This thesis, and its associated publications, help reduce this gap.

Argumentation schemes

This is the first time expert knowledge of how to interpret *in situ* gene expression information for the developmental mouse has been captured. In a series of meetings, the expert's knowledge was modelled in a number of argumentation schemes, before being translated into inferences rules for use by the argumentation engine.

During this work the range of impediments associated with such an endeavour are explored and documented. It transpires it is impossible to provide a complete picture of the expert‘s knowledge; regardless of the method used. Even if it were possible, the model would not cover the full range of reasoning within the domain as different experts reason in different ways, i.e. multiple experts need to collaborate. The schemes featured in this thesis provide a building block for future work. They can be used as a foundation for extending the current work, or they may be analysed in order to understand (or perhaps classify) the arguments used by biologists in this domain. These schemes potentially could become part of a larger corpus designed to help argumentation scholars document and study biological reasoning in general. Moreover, the schemes can be used by other systems wishing to generate natural language arguments and explanations.

Additionally, the process undertaken to capture the expert knowledge as schemes and translate it into inference rules is documented. The experience presented should enable forthcoming researchers to avoid some of the obstacles encountered during this exercise.

Architecture

In order to evaluate the ideas contained within this thesis, it was necessary to implement a prototype. Two such systems were conceived. Both followed a high-level architecture, and corresponding workflow, designed specifically for the purpose.

A key component of both the architecture and the workflow is that they are sufficiently high-level to allow the users to include whichever source(s) and argumentation toolkit they wish.

One of the main lessons learnt, when developing the prototypes, was the need for flexibility. Biological resources are heterogeneous and often difficult to integrate. When deciding whether or not to use local copies of data and caching, it is necessary to consider the availability, performance, and functionality of the individual resources being integrated. A solution may be appropriate for one resource, but inappropriate for a second. As the resources evolve over time, these decisions may need to be revised.

Presentation of arguments

The presentation of biological information and knowledge is often complex. The addition of arguments and argumentation adds another layer of difficulty. Before this work, there was no study of how arguments and argumentation could be presented to a biological user group - was there even a need to specialise the presentation?

Whilst this document has not found the ideal presentation mechanism, if such a thing exists, it has shed light onto a number of important aspects. Firstly, the results of the evaluation suggest that no single presentation mechanism is likely to satisfy the community. Secondly, the commonly used presentation styles within traditional argumentation theory are insufficient. Of the common presentation methods, a Toulmin-like left-to-right approach seems most suitable. Yet, subsequent work in Argudas implies that it too can be improved upon.

Argudas clearly shows that biologists wish to scan information, and are unwilling to read large volumes of text, regardless of whether that text is in a paragraph or in a diagrammatic form. Therefore, the ideal solution will allow both scanning of arguments and provide the support (reasoning) non-experts need. Whilst Argudas allows the former it does not supply the latter, leaving the presentation of arguments as an open question.

Application

Argumentation is not solely about reasoning. It is successfully used for natural language generation, and within pedagogical environments. In the former it can help improve communication between an expert and a non-expert, or between a system and its users. Clearly, there are many roles for such technology in almost any domain. In the classroom argumentation helps improve the scientific writing, and debating, of students. Again, this is very worthwhile; however, it is not immediately applicable to the current use case.

Focusing on this use case, using argumentation for reasoning seems profitable. As argumentation is a form of non-monotonic reasoning it can handle the incompleteness and inconsistency inherent in the use case. Computational argumentation has not been applied, for this purpose, within the biological domain before. The only previous work attempted to model expert reasoning in order to analyse the output of a tool, not to tackle issues discussed above.

In order to successfully apply argumentation it is necessary to comprehend the domain and the flow of information (and knowledge) within it. Analysing this leads to an understanding of what an argument is - a reason to believe a gene is (not) expressed - and what form the argumentation should take (reviewed in detail Section 13.2).

Argudas

Ultimately, this line of work produced a real world tool to help biologists tackle the problems of their use case. Argudas clearly demonstrates the potential of argumentation by using the metaphor of an argument to construct a platform that allows knowledgeable biologists to quickly reason for themselves using the data from a range of underlying resources. By aggregating that data, and presenting its key attributes, Argudas allows biologists to rapidly survey the field and make decisions without having to directly consult multiple resources.

Current issues

This thesis contains a substantial discussion on what is hindering the progress of computational argumentation within the life sciences. Naturally, this discussion is closely related to ideas for developing future work. Some of the central themes of these debates will be recapped below.

There are clearly issues impeding the integration of biological resources. They ensure that all knowledge systems are handicapped. Work to alleviate these concerns is ongoing; however, they will continue to have a considerable impact over both the short and medium horizons.

A second biological barrier, is the level of subjectivity within the field. Unlike medics, biologists often do not publish best practice guidelines. Argumentation seems an ideal solution to this as it allows multiple opinions to be used within a single reasoning process. Yet, capturing all these different opinions and knowing when to trust one expert more than another is not a trivial matter.

One final issue relates to the workflow undertaken for this thesis. Whilst this document clearly demonstrates computational argumentation can be applied within the life sciences, there is no doubt that the process followed here is expensive. This is largely due to the cost of capturing expert knowledge, and so is not a reflection on argumentation *per se*; however, it may reduce the willingness of others to follow the approach outlined here.

Ideas for development

As highlighted in the proceeding paragraph, the argumentation community needs to develop a range of tools and supporting methods to improve the reliability and usability of argumentation. The existing tools are often academic proof of concepts, lacking the scalability necessary for real deployment. Additionally, the range of tools needs to be increased. For example, where is the user-friendly program that helps users build a library of schemes? Moreover, why is there no published method telling non-experts how to do so?

Currently, the argumentation world assumes it is communicating with fellow scholars. However, for argumentation-centric methods to break through into the mainstream, argumentation theory has to start addressing people who do not have a traditional background. Furthermore, they must do a better job of aggregating and presenting their output to the "real world". It is too demanding to search for argumentation related activity. The whole panoply of argumentation activity should be presented via a single portal that provides instruction to users regardless of their background: expert, intermediate or beginner.

Future role

This thesis outlines a number of critical aspects for the future of argumentation within this domain. Nevertheless, the hypothesis with which this work began remains intact. Originally, argumentation was explored because it was believed that the sheer simplicity of the concept yielded a raw intuitiveness that made it amenable in ways that other forms of reasoning were not.

It appears that argumentation may, in some areas, need to be supplemented with more opaque forms of reasoning, for example Bayesian networks, yet the instinctive nature of argumentation should still carry significant potential.

13.4 Future work

In some senses this document, and the work it describes, is practically complete. Argumentation has been explored and conclusions have been drawn. As the preceding text makes clear, there is much to do before argumentation can be pragmatically applied within the life sciences. It would be simplistic to suggest, therefore, that until argumentation technology moves forward there is little to be done.

Earlier in this work a raft of potential next steps was discussed: taking the current approach into additional sub-domains of biology, developing the presentation of arguments and argumentation, improving the scalability and reliability of argumentation engines and similar toolkits, creating methods and supporting tools that can be used by non-experts to generate and maintain libraries of schemes, and the production of materials to help users better understand argumentation theory. As these ideas have been examined before, the details shall not be rehashed here.

Instead, this section shall conclude with the following remark. Unless biologists play an active role in the development of argumentation, argumentation will not evolve in a way that suits the life sciences. Likewise, unless experts from domains such as psychology and sociology are drawn in, there is the real likelihood that argumentation will never fulfil its potential as a bridge between a biologist and his/her computer. Accordingly, all future work should start with the goal of being as multidisciplinary as possible.

13.5 Final thoughts

The raw potential of argumentation in A.I. is clear. Yet, currently, one must make a distinction between the use of argumentation as a metaphor, and the application of argumentation-centric tools and methods provided by the argumentation community. The former can be successfully applied today, whilst the latter is still awaiting its time.

Appendix A

Full list of schemes

Schemes are given in two parts. Part one is the natural language inference rule; part two is the critical questions. The latter is identifiable by the use of an enumerated list; however, not all schemes have these questions. Following the scheme some comments may be provided.

Please note the following:

- Capital letters, e.g. E, indicate variables - unless it is the name of an experimental technique: ISH, or SAGE;
- The $\neg$ symbol is used to imply the opposite or negative of something - if result R indicates that a gene is expressed in a tissue, $\neg R$ indicates the gene is not expressed;
- 'is (not) expressed' should be taken to mean that the scheme can work either when the gene is expressed or is not expressed;
- The result of an experiment will vary according to the technique used, e.g. ISH produces an image showing a section of a subject, whereas microarrays provide a picture of the array itself. However, here the term *result* will be used to refer to the information published[1] indicating the researcher's belief a gene is (not) expressed in a tissue;

[1] 'Published' in the broadest sense as the information need not be printed in a journal, merely submitting the information to EMAGE for inclusion in their resource also counts.

- The term *annotation* will refer to the EMAGE/GXD Editor's interpretation of the *result*. It is the annotation which is published in their database - in the case of EMAGE this can be spatial or textual. Note the annotation may differ from the *result* on which it is based;

- *Verdict* will be used to indicate the view - gene is (not) expressed - supported by the result or annotation. Can also refer to the final conclusion reached by the user.

Please note that the term *verdict* is an artificial construction used here to simplify communication. There was a need to distinguish between experimental results and conclusions drawn from those results; however, the term 'conclusion' was not used to avoid confusion with the term given to the result of an inference rule.

Most schemes have an associated 'STRENGTH', this indicates the expert assigned degree of belief. As such, it represents the level of confidence the user has in any arguments created from the scheme. The schemes that were retracted by the expert do not have associated strengths, likewise, some that could not be implemented have no assignment.

The current section will be divided according to the subject area of the schemes.

A.1 Experimental Reliability

The schemes in this section are loosely related to experiments; however, in many cases the underlying reasoning would be equally applicable to annotations.

There are N number of experiments indicating result R
The greater the number of experiments supporting the same result, the more likely the result is to be true
Therefore R is N to be true

This scheme was retired in meeting three, as the expert felt there was no difference between 2 and more than 2 experiments with the same result. It also failed to take account of the idea that the experiments could have been performed by the same person/team that made the same mistake multiple times.

Experiment E indicates result R

Experiment $E2$ indicates result R

Two experiments indicating the same result, increases the likelihood of the result being correct

Therefore R is probably correct

1. Are E and $E2$ performed by the same lab?

2. What do other experiments/resources show?

STRENGTH=99%.

It should make no difference if $E2$ repeats the procedure of E or if $E2$ is an entirely new experiment.

Experiment E produced result R

Experiment $E2$ produced result $\neg R$

Experiment E is more likely to be accurate than $E2$

If an experiment is more likely to be accurate than a second experiment, its result is more likely to be accurate too

Therefore R is more likely to be correct than $\neg R$

1. Are E and $E2$ performed by the same lab?

2. What do other experiments/resources show?

STRENGTH=85%.

Annotation A was derived from experiment E

Annotation $\neg A$ was derived from experiment $E2$

Experiment E is more likely to be accurate than $E2$

If an experiment is more likely to be accurate than a second experiment, its annotation is more likely to be accurate too

Therefore A is more likely to be correct than $\neg A$

1. What do other experiments/resources suggest?

STRENGTH=85%.

Experiment E produced result R

E cannot be trusted

If an experiment cannot be trusted, its result cannot be trusted

Therefore R cannot be trusted

STRENGTH=99%.

Annotation A was based on experiment E

E cannot be trusted

If an experiment cannot be trusted, an annotation based on it cannot be trusted

Therefore A cannot be trusted

STRENGTH=99%.

Verdict V was based on experiment E

E cannot be trusted

If an experiment cannot be trusted, a verdict based on it cannot stand

Therefore V is withdrawn

STRENGTH = 99%.

A.1.1 Researcher Reliability

Included in this subsection are schemes that feature the researcher (or research team) that carried out the experiments.

Lab (or team leader) L has worked on gene G throughout their entire career

Lab (or team leader) $L2$ has not focused on any particular gene

L has performed experiment E on G suggesting result R

$L2$ has performed experiment $E2$ on G suggesting result $\neg R$

A lab (or team leader) who specialises on a gene is more likely to produce a correct experiment, when dealing with that gene, than a lab (or team leader) who does not specialise

Therefore E is more likely to be accurate than $E2$

1. What do other experiments/resources show?

STRENGTH=50%.

The theory behind this scheme is that a specialist will be more likely to create a fully working optimised probe, and thus perform a better experiment. The actual results should not matter - i.e. $E2$ could produce R and E suggest $\neg R$.

Here one would presume that the expert would check to see if L and $L2$ were actually experts, i.e. that a critical question to check this should exist. However, the expert has accumulated such knowledge over the years and does not need to do so. Therefore, for the purposes of recording his knowledge and practises, this question does not exist. However, for practical implementation, such a question must be employed.

Conversely, implementation did not occur. This was due to the complexity of determining whether someone was an expert on a particular area. Evidently this requires the ability to text mine over a resource such as PubMed[2], and another set of schemes.

[2] www.ncbi.nlm.nih.gov/pubmed/

Researcher team T performed experiment E

User has no (low) confidence in T

If the user has no (low) confidence in the researcher team, they will have no (low) confidence in their experiments

E is not trustworthy

STRENGTH=50%.

A.1.2 Screening programs

Schemes found in this portion of the document, all relate to screening programs (large-scale industrial research).

A screening project performed experiment E that produced result $\neg R$

A small scale research project performed experiment $E2$ that produced result R

Screening projects are less focused than optimised experiments, and do not optimise their probes as effectively

Therefore experiment $E2$ is more likely to be accurate than E

1. Is the small scale research project trustworthy?

2. What do other resources/experiments suggest?

STRENGTH = 80%.

The intuition is that the less efficient probes lead to less efficient experiments, and thus less reliable results.

Research team RT behind experiment E report that they used probe P

RT is part of a screening program

Screening programs are good at documenting their probe information

Therefore, it is very likely that E did use P

STRENGTH = 65%.

Neither this scheme nor the next one were implemented, as it is not obvious how they can be aligned to the other schemes and the central purpose of the argumentation process.

Research team RT behind experiment E report that they used probe P

RT is not part of a screening program

Normal labs are not as good at documenting their probes as screening programs

Therefore, it is likely that E did use P

STRENGTH = 80%.

A.1.3 Journal Reliability

These schemes capture the idea that journals will notice and correct most simple mistakes.

Experiment E was published in a peer reviewed journal

Experiments that have been peer reviewed in good journals are likely to be correct

E is likely to be correct

STRENGTH=15%.

Not implemented, as it is very similar to, but much weaker than, the following scheme.

Experiment E was published in a peer reviewed journal P

User has C confidence in P

The user's confidence in a journal will directly impact on experiments published in it

Confidence in E is at most C

STRENGTH=50%.

Experiment E was published in a specialised peer reviewed journal

Experiments that have been peer reviewed in specialised journals are very likely to be correct

R is very likely to be correct

In biology there are a small number of journals that are very specialised. The initial theory behind this scheme was that such a journal would be more likely to have a reviewer that really knew the field, and thus would be more likely to identify errors, for example in the naming of tissues. However, this scheme was dismissed by the expert because he rejected the theory.

A.1.4 Results

Gene G is a housekeeping gene
Housekeeping genes are of little or no interest because they are everywhere
Therefore G is of no interest

Housekeeping genes are of no interest because they rarely affect anything. They are not normally responsible for any feature (good or bad) and are ignored often.

Data is not available in EMAGE and/or GXD to find housekeeping genes, again other resources would be required.

Pattern clarity

All schemes in this section have the same questions; however, for brevity they are included only with the first scheme.

ISH Experiment E produced result R
E has low pattern clarity
Experiments with low pattern clarity are very hard to analyse and so their results are not totally trustworthy
Therefore R is possibly correct

1. What does SAGE show?
2. What does microarray show?
3. Is the equivalent result true in the previous stage?
4. Is the equivalent result true in the subsequent stage?

STRENGTH=50%.

The use of the word *possibly* in the conclusion is perhaps surprising. However, biologists often add extra qualifiers in order to emphasise the uncertainty inherent within the domain.

ISH experiment E produced result R

E has medium pattern clarity

Experiments with medium pattern clarity can be analysed with a degree of confidence, so results are trustworthy

Therefore R is likely to be correct

STRENGTH=65%.

ISH experiment E produced result R

E has high pattern clarity

Experiments with high pattern clarity are easy to analyse and so their results are trustworthy

Therefore R is very likely to be correct

STRENGTH=80%.

Image analysis

There is an implicit assumption here that a user knows how to analyse the image - yet the expert believes that is true. Indeed, he goes further, stating that the average user will be at least as good as him at analysing the image(s). As the expert did not wish to question his contemporaries, there are no critical questions for these schemes, and each scheme has a strength of 100%.

ISH experiment E suggests result R

The user, when examining the image from E, has confidence level C in R

The image is the actual result of the experiment, and so analysing it gives an accurate result

R is C likely to be true

ISH experiment E suggests that gene G is expressed in tissue T

ISH experiment $E2$ suggests that G is not expressed in T

User thinks that E and $E2$ are looking at different parts of T

User's analysis of images should take priority over editor's analysis

Therefore, G is both expressed and not expressed in T

This scheme requires expert interaction from someone able to process the two experimental images. As this could not be guaranteed, this scheme was not implemented in the final version - though it was used in the first prototype.

Cross species expression

This scheme should apply to a range of species, not just the zebrafish and mouse. Although the expert suggested that this scheme could be of use, he himself did not apply it.

Experiment E on the mouse produces result R for gene G and tissue T

Experiment $E2$ on the zebrafish produces result R for the zebrafish equivalents of G and T

Finding the same result in multiple species provides a very good indication the result is correct

Therefore R is very likely to be correct

It was not used in the system because of the sheer complexity of including it. It requires a range of other resources to provide information, and thus, theoretically, several sets of new schemes. As there was no chance of this scheme being utilised, the expert was not asked to assign a strength to it.

A.1.5 Method

Schemes relating to the experimental method are provided here.

Technique

The schemes here combine to produce a comparison between the *in situ* hybridisation (ISH), serial analysis of gene expression (SAGE), and microarray techniques. These schemes were not fully implemented, due to the extra work involved in creating a set of schemes for SAGE and another for microarray. For this reason, they were not presented to the expert to have a degree of belief assigned.

All schemes have similar questions, but for brevity they are shown only in scheme 1.

ISH experiment E produced result $\neg R$

SAGE experiment $E2$ produced result R

SAGE is more sensitive than ISH and thus may detect genes it cannot

Therefore R is probably correct

1. What does microarray suggest?

2. Can you tell what structure has been analysed?

3. Do you trust the research team?

The second and third critical questions are connected. They both relate to the idea that ISH provides a direct image of the subject (tissue experimented on), whereas SAGE (and microarray) do not. Thus when the researchers say they experimented on the eye, a user cannot be sure if they mean just the eye ball, or are including some of the optic nerve. Therefore the user cannot be sure of where the gene was actually expressed.

ISH experiment E produced result $\neg R$

microarray experiment $E2$ produced result R

microarray is more sensitive than ISH and thus may detect genes it cannot

Therefore R is probably correct

ISH experiment E produced result R

SAGE experiment $E2$ produced result $\neg R$

ISH is less sensitive that SAGE, and should not find genes that SAGE misses

Therefore neither result can be trusted without further investigation.

ISH experiment E produced result R

microarray experiment $E2$ produced result $\neg R$

ISH is less sensitive that microarray, and should not find genes that microarray misses

Therefore neither result can be trusted without further investigation.

Probes

A probe is the chemical used to detect a gene (via the protein/RNA) in a subject - normally a well designed probe will identify one gene.

Experiment E uses probe P to detect gene G

Experiment $E2$ uses P to detect gene $G2$

When the same probe is used to detect different genes that probe has been badly designed and cannot be trusted

Therefore P cannot be trusted

1. Is the probe designed to detect multiple genes?
2. What do other experiments/resources show?

The intuition is that a probe is supposed to bond with, and thus identify, a single gene. So if it can bond with two genes, it is not well designed. However, the critical question points out that a very small minority of probes are designed to detect multiple genes.

Due to the complexity and resources required to pre-compute information for this scheme it could not be implemented, accordingly it was not presented to the expert for him to assign a strength.

ISH experiment E produced an image showing gene G expressed in area A

ISH experiment $E2$ produced an image showing G expressed in area B

A does not overlap with B

E and $E2$ used different probes

Different probes may work differently, and thus produce different results

Therefore G may be expressed in both A and B

STRENGTH = 99%.

The use of the word 'may' in the conclusion of this scheme highlights the experts desire to recognise the uncertainty naturally occurring within the life sciences.

This scheme requires expert interaction by someone able to process the two experimental images. As this could not be guaranteed, this scheme was not implemented in the final version - though it was used in the first prototype.

ISH experiment E produced an image showing gene G expressed in area A

ISH experiment $E2$ produced an image showing G expressed in area B

A does not overlap with B

E and $E2$ used the same probe

The same probe should provide the same result; failure to do so indicates a fundamental problem

Therefore there is a problem with either E or $E2$ and neither can be trusted

1. What do other experiments/resources show?

STRENGTH=85%.

This scheme requires expert interaction from someone able to process the two experimental images. As this could not be guaranteed, this scheme was not implemented in the final version - though it was used in the first prototype.

The probe P for experiment E has not been sufficiently documented

If the probe has not been properly documented, the experiment cannot be repeated and verified

E cannot be fully trusted

STRENGTH=45%.

Probe P cannot be trusted

Experiment E used P to gain result R

When the probe cannot be trusted, the experiment cannot be trusted

Therefore E cannot be trusted

1. What do other experiments/resources show?

STRENGTH = 85%.

This scheme was not implemented because the schemes that concluded a probe was not reliable were not implemented.

A.2 Annotation Reliability

Textual annotation TA indicates verdict V

Spatial annotation SA indicates verdict $\neg V$

Textual annotations are more reliable than spatial annotations

Therefore V is probably correct

1. What do the other annotations/resources suggest?

STRENGTH = 99%.

Annotation A indicates verdict V for tissue T

Annotation $A2$ indicates verdict $\neg V$ for tissue $T2$

$T2$ is the parent of T

By application of the scheme 4 from Section A.3.1: $\neg V$ is true for T

Direct annotation is more likely to be correct than indirect annotation

Therefore, V is probably true for T

1. What do other annotations/resources show?

STRENGTH = 55%.

Experiment E has been analysed to produce annotation A

The user has C confidence in E

Confidence in the experiment directly impacts on the confidence of the annotation

Therefore, at most the user can have C confidence in A

STRENGTH = 99%.

The confidence in the annotation can be lower than the confidence in the experiment, e.g. the user may not trust the annotator. This seems to be implicitly handled by the weakest link algorithm, hence this was not implemented.

Annotation A suggests verdict V

The user has C confidence in A

The confidence in the annotation directly impacts on the confidence of the verdict

Therefore, at most the user can have C confidence in V

STRENGTH = 99%.

This seems to be implicitly handled by the weakest link algorithm, hence this was not implemented.

Experiment E, in EMAGE, has annotation A

Experiment E, in GXD, has annotation $\neg A$

Because the editors are more experienced, EMAGE is more reliable than GXD

Therefore A is probably correct

1. Is E the only relevant experiment in EMAGE?
2. Is $E2$ the only relevant experiment in GXD?
3. What do the other resources suggest?

For an experiment to be 'relevant' it should consider the same gene and tissue.

This scheme seemed very biased, so instead of implementing it, the following scheme was utilised instead.

Experiment E, in EMAGE, has annotation A

Experiment E, in GXD, has annotation $\neg A$

A difference in annotation between EMAGE and GXD is a reason to doubt both annotations

Therefore neither A nor $\neg A$ can be trusted

1. Is E the only relevant experiment in EMAGE?
2. Is $E2$ the only relevant experiment in GXD?
3. What do the other resources suggest?

STRENGTH=85%.

Verdict V was based on annotation A

A cannot be trusted

If an annotation cannot be trusted, a verdict based on it cannot stand

Therefore V is withdrawn

STRENGTH = 99%.

The idea captured by this scheme seems implicit in argumentation, thus this scheme was not implemented.

Annotation A indicates verdict V

Annotation $A2$ indicates verdict $\neg V$

Annotation A is more likely to be accurate than $A2$

If an annotation is more likely to be accurate than a second annotation, its verdict is more likely to be accurate too

Therefore V is more likely to be correct than $\neg V$

1. What do other annotations/resources show?

STRENGTH = 85%.

The idea captured by this scheme seems implicit in argumentation, thus this scheme was not implemented.

Annotation A for ISH experiment E, indicates verdict V

The pattern clarity of E is low

If the pattern clarity is low, annotations are almost useless

Therefore, A does not provide a reason to believe V

STRENGTH = 50%.

This scheme is not implemented as it is very similar to another scheme that has been implemented.

A.2.1 Spatial

The schemes for spatial annotations should apply equally to spatial annotations and textual annotations derived from spatial annotations.

Spatial annotation SA suggests verdict V

Spatial annotations are a reasonable indication of expression

Therefore it is likely that V is true

1. Is SA reliable (correct)?
2. What do the textual annotations show?
3. How good is the morphological match?
4. How high is the pattern clarity?

STRENGTH=65%.

ISH experiment E produced result R, which was converted into a spatial annotation SA by EMAGE

The research team behind E approved (created) SA

Spatial annotations performed by, or approved by, the research team are very likely to be correct

Therefore SA is very likely to be correct

1. What do the textual annotations show?
2. Are the research team trustworthy?
3. How good is the morphological match?
4. How high is the pattern clarity?

STRENGTH = 1%.

This basic scheme can be adapted to the situation where the research team perform the spatial annotation rather than just approve it. The expert is making an assumption that screening programs do not create or approve mappings (due to the large volume of data they deal with). As those programs are often less expert than the editors at EMAGE, if they ever were to be involved with a spatial annotation, confidence in that annotation should fall. Again, this should be tested with a question - but the expert does not do so.

This scheme was not implemented as the strength was too low for it to be worthwhile.

The spatial annotation SA for experiment E indicates verdict V

SA is based on spatial mapping SM

SM features N number of voxels

$N < 10$

Spatial mappings with less than 10 voxels are likely to be errors in the mapping process

SA is not a good indicator that V is true

This scheme was removed in the third meeting because it was too simplistic to be accurate; however, the reality was too complex (or not well enough researched) to model.

The spatial annotation SA for experiment E indicates verdict V

SA is based on spatial mapping SM

SM features N number of voxels

$10 < N < 50$

Spatial mappings with more than 10, but less than 50, voxels indicate the result may be accurate

SA is an indicator that V is true

Removed, for the same reason as the previous scheme.

The spatial annotation SA for experiment E indicates verdict V

SA is based on spatial mapping SM

SM features N number of voxels

$50 < N$

Spatial mappings with more than 50 voxels indicate the result is very likely to be accurate

SA is a good indicator that V is true

Removed, for the same reason as the previous scheme.

Spatial annotation SA suggests gene G is expressed in X% of tissue T

Spatial annotation $SA2$ suggests G is not expressed in Y% of T

$Y < X$

The greater the % the more meaningful the spatial annotation

Therefore SA is probably the more reliable annotation

1. What do the textual annotations show?
2. How do the morphological mapping scores compare?
3. How do the pattern clarity scores compare?

The intuition works regardless of whether X indicates the gene is, or is not, expressed.

STRENGTH = 50%

Spatial annotation SA suggests that gene G is (not) expressed in X% of tissue T

$X < 3$%

A spatial annotation of less than 3% is likely to be stray voxels as a result of a bad mapping, and can be ignored

SA can be ignored (SA does not provide a reason to believe G is (not) expressed in T)

1. What do the textual annotations show?
2. What is the morphological mapping score?
3. What is the pattern clarity score?

STRENGTH = 80%.

Spatial annotation SA was derived from ISH experiment E

SA was derived from spatial mapping SM

SM was performed by individual I

The user does not trust I to perform spatial mappings

If the individual performing the mapping is not trusted, the mapping cannot be trusted, and nor can the annotation based on it

SA cannot be trusted

STRENGTH = 99%.

No questions were asked because the expert felt it unnecessary to question the opinions of fellow domain experts.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

The user has more trust in RT than $RT2$

Spatial annotation SA was derived from E

Spatial annotation $SA2$ was derived from $E2$

SA indicates V, but $SA2$ indicates $\neg V$

When 2 spatial annotations disagree we can resolve the dispute by choosing the annotation from the experiment that was performed by the research team the user has more trust in

V is more likely to be correct

1. How do the morphological mapping scores compare?

2. How do the pattern clarity scores compare?

STRENGTH = 55%.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

The morphological match for E is greater than $E2$

Spatial annotation SA was derived from E

Spatial annotation $SA2$ was derived from $E2$

SA indicates V, but $SA2$ indicates $\neg V$

When 2 spatial annotations disagree we can resolve the dispute by choosing the annotation from the experiment with the higher morphological match

V is more likely to be correct

STRENGTH = 99%.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

The pattern clarity is higher for E than $E2$

Annotation A was derived from E

Annotation $A2$ was derived from $E2$

A indicates V, but $A2$ indicates $\neg V$

When 2 annotations disagree we can resolve the dispute by choosing the annotation from the experiment with the higher pattern clarity score

V is more likely to be correct

1. If A and $A2$ are from different resources, which resource does the user have most trust in?

2. Which researcher does the user have most trust in?

STRENGTH = 85%.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

Spatial annotation SA was derived from E by mapper M

Spatial annotation $SA2$ was derived from $E2$ by mapper $M2$

SA indicates V, but $SA2$ indicates $\neg V$

The user has more trust in M than $M2$

When 2 spatial annotations disagree we can resolve the dispute by choosing the annotation by the mapper the user has more trust in

V is more likely to be correct

STRENGTH = 95%.

The morphological match and pattern clarity are not referred to here, as they are both assigned by the mapper.

Morphological match

Spatial annotation SA was derived from ISH experiment E

The morphological match for E is low

If the morphological match is low, there is a low chance of the mapping being spatially accurate

SA is unlikely to be accurate

1. What is the pattern clarity?

STRENGTH = 50%.

Spatial annotation SA was derived from ISH experiment E

The morphological match for E is medium

If the morphological match is medium, there is a reasonable chance of the mapping being spatially accurate

SA is possibly accurate

1. What is the pattern clarity?

STRENGTH = 65%.

Spatial annotation SA was derived from ISH experiment E

The morphological match for E is high

If the morphological match is high, there is a good chance of the mapping being spatially accurate

SA is likely to be accurate

1. What is the pattern clarity?

STRENGTH = 80%.

A.2.2 Textual

Textual annotations are created by the database editor, and require them to map the result generated from the experiment to their own representation (EMAP anatomy). Note: the schemes here do not apply to textual annotations derived from spatial annotations.

ISH experiment E produced result R, which was converted into a textual annotation TA by EMAGE

Textual mappings are very regularly correct

Therefore TA is very likely to be correct

1. Is the experiment reliable?

2. What do other resources show?

STRENGTH=80%.

Textual annotation TA was derived from ISH experiment E

TA was created by individual I

The user does not trust I to perform textual annotations

If the individual creating the annotation is not trusted, then the annotation cannot be trusted

TA cannot be trusted

STRENGTH = 99%.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

Textual annotation TA was derived from E by indexor I

Textual annotation $TA2$ was derived from $E2$ by indexor $I2$

TA indicates V, but $T2$ indicates $\neg V$

The user has more trust in I than $I2$

When 2 textual annotations disagree we can resolve the dispute by choosing the annotation that was performed by the indexor the user has more trust in

V is more likely to be correct

STRENGTH = 65%.

Experiment E was performed by lab RT

Experiment $E2$ was performed by lab $RT2$

The user has more trust in RT than $RT2$

Textual annotation TA was derived from E

Textual annotation $TA2$ was derived from $E2$

TA indicates V, but $TA2$ indicates $\neg V$

When 2 textual annotations disagree we can resolve the dispute by choosing the annotation from the experiment that was performed by the research team the user has more trust in

V is more likely to be correct

1. How do the morphological mapping scores compare?
2. How do the pattern clarity scores compare?

STRENGTH = 50%.

A.3 Anatomical Location

The schemes contained in this subsection all rely on the notion that a tissue often develops, yet still persists, from one Theiler Stage to the next. Furthermore, they all highlight the fact that a gene's expression level will usually persist from one stage to the next.

Experiment E produced result R for gene G and tissue T

Experiment $E2$ produced R for G and tissue $T + 1$

$T + 1$ is the equivalent of T but in the next Theiler Stage

The same result in two adjacent stages indicates the result is likely to be true

Therefore R is likely to be true

1. Are the experiments reliable?
2. What do other resources indicate?

STRENGTH = 85%.

The reasoning behind this scheme also works for T-1 (the same tissue in the previous Theiler Stage).

Experiment E produced result R for gene G and tissue T

Experiment $E2$ produced R for G and tissue $T+1$

Experiment $E3$ produced R for G and tissue $T-1$

$T+1$ is the equivalent of T but in the next Theiler Stage

$T-1$ is the equivalent of T but in the previous Theiler Stage

The same result in previous and next stage indicates that the result is extremely likely to be true in the current stage

Therefore R is extremely likely to be true

1. Are the experiments reliable?
2. What do other resources indicate?

STRENGTH = 85%.

This scheme is essentially the same as the previous scheme, and so was not implemented.

There is no experiment E suggesting R for gene G in tissue T

Experiment $E2$ suggests result R for G in tissue $T-1$

Experiment $E3$ suggests result R for G in tissue $T+1$

$T-1$ is the equivalent of T but in the previous Theiler Stage

$T+1$ is the equivalent of T but in the next Theiler Stage

When the same result occurs in the stage before, and the stage after, it is extremely likely to be true for the stage in the middle

R is almost certainly true for G in T

1. Are the experiments reliable?
2. What do other resources indicate?

STRENGTH = 85%.

Gene G appears in low levels in almost every tissue
Genes that appear everywhere are housekeeping genes
Therefore G is a housekeeping gene

This scheme was removed as the expert felt it would be more likely to be the subject of a set of SAGE schemes as it would not be possible to determine housekeeping genes using EMAGE or GXD.

A.3.1 Ontological rules

The schemes in this subsection all exist because the EMAP anatomy uses part-of relationships between the tissues (nodes) in the graph.

Experiment E suggests gene G is expressed in tissue T
T is a component of tissue $T2$
If a gene is expressed in a component, it must be expressed in the parent
Therefore G is expressed in $S2$

STRENGTH = 100%.

Experiment E suggests gene G is not expressed in tissue T
T is a component of $T2$
If a gene is not expressed in a component, it may be expressed in a different child of the parent
Nothing can be inferred about the relationship between G and $T2$

STRENGTH = 15%.

This scheme captures the knowledge that a user cannot propagate positive gene expression down through the EMAP anatomy.

This scheme is very difficult to implement, and its low strength makes it pointless to do so, hence it was not implemented.

Experiment E suggests gene G is expressed in tissue T

T is a parent of $T2$

If a gene is expressed in a parent, it may be expressed in a different child

Nothing can be inferred about the relationship between G and $T2$

STRENGTH = 65%.

Experiment E suggests gene G is not expressed in tissue T

T is a parent of $T2$

If a gene is annotated as being 'not expressed' in a parent, it cannot be expressed in a child

Therefore G is not expressed in $T2$

STRENGTH = 100%

Tissue T has gene G (not) expressed in it

Tissue $T2$ is the child of T

$T2$ is T's only child

When a tissue has only 1 child, the child and parent are identical

Therefore, G is (not) expressed in $T2$

This resolves an error in the EMAP anatomy. A second scheme, for upwards propagation (passing expression to parent) can also exist but is not shown.

Because this error was so unusual, this scheme was not implemented, nor shown to the expert for a strength to be assigned.

Appendix B

Inference rules

This chapter contains the inferences rules derived from the schemes in Chapter 6 according to the method in Chapter 7.

```
r2de(E) <- researcher(E, RT), trust(RT, 'low') 0.5.
r2de(E) <- publisher(E, J), trust(J, 'low') 0.5.
r2de(E) <- \+(probe(E, P)) 0.45.
r2te(E) <- publisher(E, J), trust(E, 'med') 0.5.
r2te(E) <- publisher(E, J), trust(E, 'hi') 0.5.

betterexperiment(E, E2) <- \+(screen(E)), screen(E2) 0.8.

sa(E, G, T, 'expressed', P, indirect) <- sa(E, G, T2, 'expressed', P, direct),
            component(T2, T).
sa(E, G, T, 'absent', P, indirect) <- sa(E, G, T2, 'absent', P, direct),
            component(T, T2).

betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
           \ ==(L, L2), >(P1, P2) 0.5.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
           \ ==L, L2), researcher(E, RT), researcher(E2, RT2), trust(RT, 'hi'),
            trust(RT2, 'med') 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
            \ ==(L, L2), researcher(E, RT), researcher(E2, RT2), trust(RT, 'hi'),
                trust(RT2, ’low’) 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), researcher(E, RT), researcher(E2, RT2), trust(RT, ’med’),
                trust(RT2, ’low’) 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), morphmatch(E, ’hi’), morphmatch(E2, ’med’) 0.99.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), morphmatch(E, ’hi’), morphmatch(E2, ’low’) 0.99
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), morphmatch(E, ’med’), morphmatch(E2, ’low’) 0.99.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), clarity(E, ’hi’), clarity(E2, ’med’) 0.85.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
               \ ==(L, L2), clarity(E, ’hi’), clarity(E2, ’low’) 0.85.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), clarity(E, ’med’), clarity(E2, ’low’) 0.85.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), spatialmapper(E, RT), spatialmapper(E2, RT2),
                \ ==(RT, RT2), trust(RT, ’hi’), trust(RT2, ’med’) 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), spatialmapper(E, RT), spatialmapper(E2, RT2),
                \ ==(RT, RT2), trust(RT, ’hi’), trust(RT2, ’low’) 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D), sa(E2, G, T, L2, P2, D2),
                \ ==(L, L2), spatialmapper(E, RT), spatialmapper(E2, RT2),
                \ ==(RT, RT2), trust(RT, ’med’), trust(RT2, ’low’) 0.55.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, D),
                sa(E2, G, T, L2, P2, D2), \ ==(L, L2), >(P1, P2) 0.50.
betterspatial(E, G, T, E2) <- sa(E, G, T, L, P1, direct),
                sa(E2, G, T, L2, P2, indirect) 0.55.

r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), r2de(E) 0.99.
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), clarity(E, ’low’) 0.5.
```

```
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), >(3.0, P) 0.8.
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), spatialmapper(E, RT),
            trust(RT, 'low') 0.99.
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), morphmatch(E, 'low') 0.5.
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), sa(E2, G, T, L2, P2, D2),
            \ ==(L, L2), betterexperiment(E2, E) 0.85.
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), sa(E2, G, T, L2, P2, D2),
            \ ==(L, L2), betterspatial(E2, G, T, E).
r2dsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), ta(E2, G, T, L2, S, D2),
            \ ==(L, L2) 0.99.


r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), sa(E2, G, T, L, P2, D2),
            \ ==(E, E2) 0.99.
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), clarity(E, 'med') 0.65.
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), clarity(E, 'hi') 0.8.
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), morphmatch(E, 'med') 0.65.
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), morphmatch(E, 'hi') 0.8.
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), r2te(E).
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), sa(E2, G, T, L2, P2, D2),
            \ ==(L, L2), betterspatial(E, G, T, E2).
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), sa(E2, G, T, L2, P2, D2),
            \ ==(L, L2), betterexperiment(E, E2).
r2tsa(E, G, T, L, P) <- sa(E, G, T, L, P, D1), ta(E2, G, T, L, S, D2) 0.99.
r2tsa(E, G, T, expressed, P) <- sa(E, G, T, 'expressed', P, D1),
            \+(sa(E2, G, T, 'absent', P2, D2)),
            \+(ta(E3, G, T, 'absent', S, D3)).
r2tsa(E, G, T, absent, P) <- sa(E, G, T, 'absent', P, D1),
            \+(sa(E2, G, T, 'expressed', P2, D2)),
            \+(ta(E3, G, T, 'expressed', S, D3)).


ta(E, G, T, 'expressed', S, indirect) <- ta(E, G, T2, 'expressed', S, direct),
            component(T2, T).
```

```
ta(E, G, T, ’absent’, S, indirect) <- ta(E, G, T2, ’absent’, S, direct),
            component(T, T2).
ta(E, G, T, L, S, D1) <- ta(E, G, T1, L, S, D1), ta(E2, G, T2, L, S2, D2),
            lineage(T1, T), lineage(T, T2) 0.85.

bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), textualmapper(E, T1), textualmapper(E, T2),
           trust(T1, ’hi’), trust(T2, ’med’) 0.65.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), textualmapper(E, T1), textualmapper(E, T2),
            trust(T1, ’hi’), trust(T2, ’low’) 0.65.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), textualmapper(E, T1), textualmapper(E, T2),
            trust(T1, ’med’), trust(T2, ’low’) 0.65.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), researcher(E, T1), researcher(E, T2),
            trust(T1, ’hi’), trust(T2, ’med’) 0.55.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), researcher(E, T1), researcher(E, T2),
            trust(T1, ’hi’), trust(T2, ’low’) 0.55.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), researcher(E, T1), researcher(E, T2),
            trust(T1, ’med’), trust(T2, ’low’) 0.55.
bettertextual(E, G, T, E2) <- ta(E, G, T, L, S1, direct),
            ta(E2, G, T, L2, S2, indirect) 0.55.

r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), betterexperiment(E2, E) 0.85.
r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \==(L, L2), bettertextual(E2, G, T, E).
r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), r2de(E) 0.99.
r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), clarity(E, ’low’) 0.5.
```

```
r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), textualmapper(E, RT),
            trust(RT, 'low') 0.99.
r2dta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L, S2, D2),
            \ ==(L, L2), \ ==(S, S2), pubmed(E, PID1),
            pubmed(E2, PID2), \ ==(PID1, PID2) 0.85.

r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L, S2, D2),
            \ ==(E, E2), \+(same(E, E2)) 0.99.
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), clarity(E, 'med') 0.65.
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), clarity(E, 'hi') 0.8.
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \ ==(L, L2), bettertextual(E, G, T, E2).
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T, L2, S2, D2),
            \ ==(L, L2), betterexperiment(E, E2) 0.85.
r2tta(E, G, T, expressed, S) <- ta(E, G, T, 'expressed', P, D1),
            \+(ta(E2, G, T, 'absent', S2, D2)) 0.8.
r2tta(E, G, T, absent, S) <- ta(E, G, T, 'expressed', P, D1),
            \+(ta(E2, G, T, 'expressed', S2, D2)) 0.8.
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T2, L, S2, D2),
            lineage(T, T2) 0.85.
r2tta(E, G, T, L, S) <- ta(E, G, T, L, S, D1), ta(E2, G, T2, L, S2, D2),
            lineage(T2, T) 0.85.

verdict(G, T, L) <- r2tsa(E, G, T, L, P).
verdict(G, T, L) <- r2tta(E, G, T, L, S).
~verdict(G, T, L) <- r2dsa(E, G, T, L, P).
~verdict(G, T, L) <- r2dta(E, G, T, L, S).

expressed(G, T) <- verdict(G, T, 'expressed').
expressed(G, T) <- verdict(G, T, 'absent').
```

Appendix C

ASPIC goals

The following pages are a reproduction of one page from the argumentation services platform with integrated components (ASPIC) website - formerly www.argumentation.org. This page presents a brief introduction to argumentation, and finally a short summary of the goals of the project.

The goals of the project are elementary:

1. To bring together a number of research groups who have established themselves in formal argumentation theory, in order to develop consensus theoretical models that improve on and extend existing models;

2. To develop efficient proof procedures and software component implementations of these models for deployment in real-world applications.

It is the latter goal that resulted in the creation of the argumentation toolkit utilised in this work.

ASPIC

PROJECT

OVERVIEW

ASPIC DELIVERABLES

ASPIC SOFTWARE

ASPIC CONFERENCE COMMA '06

ASPIC NETWORK

IMPRINT

PARTNERS

Singular Logic S.A.

Cancer Research UK

Zeus Consulting S.A.

University of Ljubljana

Technical University of Catalonia

Institut de Recherche en Informatique de Toulouse

University of Liverpool

Utrecht University

City University of New York

LINKS

ARGUMENTATION

1. Introduction to Argumentation

The theory of argumentation is a rich, interdisciplinary area of research straddling philosophy, communication studies, linguistics, psychology and artificial intelligence. Traditionally, he focus has been on "informal" studies of argumentation and its role in natural human reasoning and dialogue. More formal logical accounts of argumentation have recently been proposed by the artificial intelligence community as a promising paradigm for modelling common-sense reasoning and communication between reasoning entities. In these works, an **argument** is a set of premises offered in support of a claim. For example, consider the argument:

A1 = Information ***I*** about Tony should be published
    because
    Tony has political responsibilities
    and
    I is in the national interest
    and
    if a person has political responsibilities and information about that person is in
the
    national interest then that information about the person should be published

where Information I about Tony should be published is the argument's claim, and the italicised statements following 'because' are the the argument's premises. Consider the following counter-argument to A1 that undermines a premise in A1:

A2 = Tony does not have political responsibilities because Tony resigned from parliament and if a person resigns from parliament then that person no longer has political responsibilities

Consider the following counter-argument to A2:

A3 = Tony does have political responsibilities because Tony is now middle east envoy and if a person is a middle east envoy then that person has political responsibilities

Argumentation is the the process whereby arguments are constructed and evaluated in light of their interactions with other arguments. So, in the above example, arguments A1, A2 and A3 have been constructed. A3 'attacks' A2 by contradicting A2's claim, and A2 'attacks' A1 by contradicting a premise in A1 (and so undermining A1). The winning arguments can then be evaluated. A1 is attacked by A2, but since A2 is itself attacked by A3, and the latter is not attacked, we obtain that A1 and A3 are the winning arguments.

This example illustrates the modular nature of argumentation that most formal theories (models) of argumentation adopt: 1) arguments are constructed in some underlying logic that manipulates statements about the world; 2) interactions between arguments are defined; 3) given the network of interacting arguments, the winning arguments are evaluated.

The appeal of the argumentation paradigm resides in this intuitive modular characterisation that is akin to human modes of reasoning. Also, recent work in the AI, and computer science community at large, has illustrated the potential for tractable implementations of logical models of argumentation, and the wide range of application of these implementations in software systems.

Furthermore, the inherently dialectical nature of argumentation models provide principled ways in which to structure exchange of, and reasoning about, justifications/arguments for proposals and or statements between human and or automated reasoning entities (agents).

Consider the above example where instead of a single agent engaging in its own internal argumentation to arrive at a conclusion, we now have two agents involved in a dialogue. Greg proposes argument A1, Alistair argument A2, and then Greg counters with argument A3. This represents a dialogue where each participant has the goal of persuading the other to adopt a belief through the process of exchanging arguments that must interact according to the underlying model of argumentation, and where the winning arguments are evaluated according to the underlying model of argumentation.

Of course, dialogues introduce an added dimension, in the sense that realistic dialogues often involve more than simply the exchange of arguments. For example, Alistair might challenge a premise in argument A1, by asking why information I is private. The burden of proof is on Greg to provide an argument as to why I is private. Otherwise, Alistair can be legitimately be said to be 'winning' the argument or dialogue. The formal study of dialogue models therefore accounts for a broader range of statements or 'locutions' than simply those involving submission of arguments.

Furthermore, the goal of the argumentation based dialogue may not only be to persuade, but also to collaboratively decide or deliberate over an appropriate course of action, or to negotiate over resources. In these dialogues, the reasons or arguments for proposed actions, or offers and rejections, can be usefully used to further the goal of the dialogue.

The goal of the dialogue may determine a specific set of statements or allowed locutions, as well as rules for making locutions at any point in the dialogue, and rules for determining the outcome of the dialogue. These rules are encoded in a dialogue's protocol.

Consider for example the following negotiation dialogue between a buyer and seller of cars in which locutions also involve making, accepting and rejecting offers:

Seller - Offer: Renault
Buyer - Reject: Renault
Seller - Why
Buyer - Argue: Because Renault is a French make of car, and french cars are unsafe
Seller - Argue: Renaults are not unsafe as Renaults have been given the award of safest car in Europe by the European Union.
Buyer - Accept: Renault

The above example illustrates the utility of argumentation based models of reasoning and their application to dialogues. Online negotiations involving automated software agents are a key area of research and development. In a handshaking protocol, a seller would simply successively make offers and have these either rejected or accepted. The exchange of arguments provides for agreements that would not be reached in simple handshaking protocols. In the above example, it is by eliciting the reason for the rejection, and successfully countering this reason, that the seller is then able to convince the buyer to buy the car.

2. ASPIC and Argumentation: The aims of the ASPIC project

ASPIC is a collaborative project involving AI researchers and computer scientists in both the academic and commercial sectors. The aim of this project is two-fold:

1) To bring together a number of research groups who have established themselves in formal argumentation theory, in order to develop consensus theoretical models that improve on and extend existing models;

2) To develop efficient proof procedures and software component implementations of

these models for deployment in real-world applications.

Appendix D

Argumentation Engine

A review of the argumentation engine used in this work now follows. A brief summary of what an argumentation framework is composed of will be conducted in Section D.1, subsequently each of the individual aspects of the framework will be explored with particular attention paid to how the ASPIC argumentation engine implements them. Section D.2 provides an analysis of the actual implementation. The chapter concludes with Section D.3 providing an example of argumentation.

D.1 Implementation of an argumentation framework

Argumentation frameworks specify an abstract "template" that when implemented allows computers to conduct argumentation automatically. Many different styles of framework exist; Prakken [42] identified five common traits: an underlying logical language; the concept of an argument; the concept of attack between arguments; the concept of defeat; and a definition of argument status (how to assess if an argument is justified). Each layer builds upon the previous one, as summarised in Figure D.1.

The following sections will examine each level in turn.

D.1.1 Logic

Prakken's description of an argumentation framework begins with the specification of a logical language on which the rest of the system is developed.

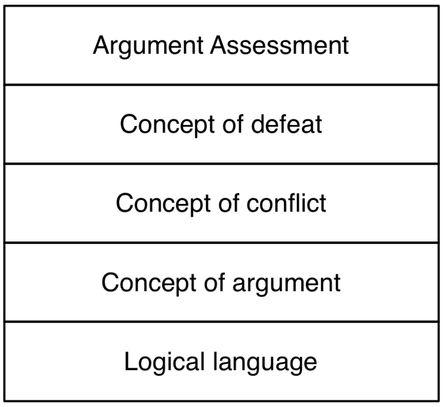


Figure D.1: Prakken's argumentation framework from [42].

An important point to note is that there are effectively three different syntaxes for anyone wishing to use the ASPIC engine. Firstly, there is the syntax used by the authors of the theory, e.g. [156]. Secondly, there is the PROLOG-like syntax that the engine uses, and finally there is the JAVA API for the engine. The JAVA API will be discussed in Section D.2.3.

Although there is a mapping between the theoretical syntax and the implemented syntax, these are not identical. For example, the theoretical syntax resembles classical logic, whereas the implemented syntax is closer to PROLOG. To illustrate the difference between the syntaxes consider the following very simple rule:

Natural language It is raining implies you may get wet

Theoretical raining $\Rightarrow$ wet

PROLOG-like wet <- raining 0.7

The use of the word *may* indicates that the rule is defeasible, i.e. can be defeated. A rule that cannot be defeated would be *strict*. In the theoretical version $\Rightarrow$ is used to indicate the defeasible nature of the rule (a strict rule would use $\rightarrow$). The implemented version uses the same arrow for both strict and defeasible rules. It is the provision of a number, following the rule, that indicates the rule is defeasible. Despite this number appearing to be a probability, the creators describe it as a *degree of belief* and as such it documents the strength of the author's confidence in that rule. The theory suggests the use of such a system is necessary, but does not specify it.

A further difference between the two syntaxes is the symbols used to indicate weak negation. Both syntaxes use $\sim$ for strong negation; however, for weak negation the theoretical syntax uses $-$ and the implemented syntax \+.

The final main difference emerges because the implemented syntax includes predefined common functions. For example, testing if one integer is greater than a second has the predicate >(X,Y) where X and Y are variables to be instantiated at run time.

Crucially, the theoretical syntax cannot be understood by the engine therefore it shall not be used hereafter.

The implemented logic

To initiate the process note the following basic terminology:

fact a piece of information;

rule an inference rule, which uses known facts to generate new facts;

degree of belief a numerical value (between 0 and 1) the author assigns to a rule or fact to indicate his/her confidence in that rule/fact;

strict a rule or fact that must be true, i.e. have a degree of belief of 1;

defeasible a rule or fact that may be true - the likelihood of this event being determined by the degree of belief.

Facts can be simple literals such as a or more complex predicates like expressed(bmp4, limb) - which states that the gene *bmp4* is expressed in the structure called limb. Notice that the individual terms are all lowercase, this indicates that they are constants. Variables start with an uppercase letter.

A simple rule to state that a gene is expressed in a particular structure because an unnamed resource says so, could be:

```
expressed(Gene, Structure) <- resourceExpressed(Gene, Structure).
```

The logic provides both weak and strong negation. Strong negation implements the idea that a fact can be true or false. False is indicated by $\sim$. Thus to suggest a gene is not expressed if the resources indicates that is so, the following could be used:

```
~expressed(G, S) <- resourceNotExpressed(G, S).
```

Weak negation captures the idea of the closed world assumption - if something is not known to be true, then it must be false. Alternatively, if something is not known to be false, then it must be true. For example, if a resource suggests a gene is expressed in a structure, and a biologist is not known to disagree, then the gene must indeed be expressed:

```
expressed(G,S) <- resourceExpressed(G,S),
                      \+(biologistNotExpressed(G,S)).
```

D.1.2 Arguments

Following on from the logic, the second layer of Prakken's framework is the notion of an argument. The concept of 'argument' was discussed earlier in Section 2.4.2, and this current discussion may be considered complimentary to it. To recap that earlier discussion, recall:

- an argument is a reason to believe something is true;
- arguments in logical fields often consist of a series of propositions that when true imply a final proposition is also true;
- many arguments in the real world have this basic structure, which corresponds to the logical notion of *modus ponens*;
- argumentation schemes, and associated critical questions, are used to capture the form of common classes of argument.

In [42] Prakken remarks that the definition of the argument is essentially the same as a proof in the logical language. One important distinction is [74]:

> arguments have the form of logical proof, but they do not have the force
> of logical proof (page 1)

This is a reminder that traditionally logical proofs are deductive, and thus show something is definitely true, whereas arguments are commonly defeasible and accordingly merely suggest something may be true.

```
Conclusion
    Premise 1
        Premise 3
        Rule 3: Premise 1 <- Premise 3
    Premise 2
    Rule 1: Conclusion <- Premise 1, Premise 2
```

Figure D.2: Layout of an argument in ASPIC's argumentation engine.

Implementation of arguments in ASPIC

The implemented structure of an argument is related closely to the chosen logic - an argument is effectively a proof in that logic.

The structure of a typical argument in ASPIC can be seen in Figure D.2. Notice the tree-like structure, in which sub-arguments are on lower levels. Also significant is that rules have associated names, e.g. Rule 1. This is because the engine assigns each rule a name, unless the user has already done so. The rules are explicitly given in the argument, therefore ensuring the argument has a *modus ponens* form.

D.1.3 Attack

Although the theory behind ASPIC distinguishes between rebuttal and undercut, the engine implements both using rebuttal.

Rebuttal works as one might expect, by having two arguments with opposite conclusions. It is possible to rebut premises, assumptions and sub-arguments too.

To undercut an inference captured in a particular rule, it is necessary to create a second rule that rebuts the name of the first rule. For example, to undercut Rule 1 (Conclusion <- Premise 1 & Premise 2) from Figure D.2, the following rule could be used: ∼Rule 1 <- Premise 4. If Premise 4 is ever true, Rule 1 cannot be applied.

This works because the name of the rule utilised is an implicit premise of an argument. Thus the argument from Figure D.2 should actually be the presentation from Figure D.3

```
Conclusion
    Premise 1
        Premise 3
        Rule 3
        Rule 3: Premise 1 <- Premise 3
    Premise 2
    Rule 1
    Rule 1: Conclusion <- Premise 1 & Premise 2
```

Figure D.3: The structure of an ASPIC argument, with the implicit premise used in undercut made explicit.

D.1.4 Defeat

The penultimate layer of an argument framework deals with conflict resolution - deciding which argument wins an individual attack.

Within the ASPIC argumentation engine, each fact and rule can be assigned a degree of belief. If this is not done, or if the user assigns a belief of 1, this results in the fact/rule being deemed strict. To propagate and aggregate each of the individual degrees of belief to the final argument, the user can choose to use weakest link or last link. Once the strength of an argument has been ascertained the following basic principles can be applied:

- because undercut is implemented as rebuttal, it has the same properties as rebuttal;
- strict arguments cannot be defeated by rebuttal - unless an assumption is rebutted (rebutting an assumption always works);
  - strict arguments defeat rebutting defeasible arguments;
- when two strict arguments rebut each other, they are both true;
- when two defeasible arguments rebut each other, the strongest argument (the argument with the highest degree of belief) wins;

- when two defeasible arguments, with the same strength, rebut each other the outcome depends on the semantics chosen for level five of Prakken's framework (they are both false under the sceptical semantics, and both true with the credulous semantics).

D.1.5 Evaluation of arguments

Throughout *evaluation* is treated purely as the process of determining if an argument is acceptable, which ties in with the notion of the fifth layer of the argumentation framework.

The ASPIC argumentation engine implements two of Dung's semantics: grounded; and, preferred credulous - discussed in Section 2.4.6.

These are viewed as a dialogue game between two players [157]. It occurs when the user submits a query - a fact that (s)he wishes the system to argue over the truth of. The query, and thus the fact it represents, can be *undefeated* (true with respect to current knowledge), *defeated* (false with respect to current knowledge), or *unknown*.

The game features two computer players, the *proponent* (PRO) who attempts to prove the query, and the *opponent* (OPP) who tries to stop PRO. The game starts with PRO creating an argument to support the query (an argument whose conclusion is identical to the query). This process begins by searching for a rule with an appropriate conclusion. Once found, rules with conclusions that are identical to the premises are sought. If the premises cannot be satisfied in this manner, the facts are examined to determine if they satisfy the premises.

OPP now attempts to defeat PRO's argument. To succeed, OPP's argument must rebut part of PRO's and have a higher degree of belief - assuming credulous semantics. OPP starts by trying to construct arguments that rebut the conclusion of PRO's argument. If that cannot be done, OPP attempts to rebut the premises - this includes performing undercuts.

If OPP succeeds in defeating PRO's argument, PRO will attempt to counter OPP's argument by defeating it. This process of attack and counter-attack continues until one player (PRO or OPP) fails to defeat the other's argument. If PRO is stopped, it tries a new line of defence by creating a new argument, to support the conclusion that OPP has defeated. If PRO fails to do so, OPP wins. However, if OPP fails, it tries to defeat one of PRO's previous arguments, if it cannot do so PRO wins.

The above is an obvious simplification, as there are two semantics (and games), one sceptical and one credulous. However, the precise details are not necessary to gain an understanding of this work.

D.2 ASPIC argumentation engine

In the previous section, the theory behind argumentation frameworks was explored, paying particular attention to how that theory related to the ASPIC argumentation engine employed in this work. This section will explore the ASPIC engine further, examining aspects of its implementation and details that are unique to it.

D.2.1 Unique additions to theory

As Caminada and Amgoud [156] point out, there are two problems prevalent in many existing theories. Namely that they do not force the list of justified arguments to be consistent or complete. Such issues can be demonstrated easily using the ASPIC engine. For example, if two strict arguments rebut each other, neither will defeat the other, and the resulting set of justified arguments will contain an inconsistency.

As a possible solution to these problems [156] specifies two additional operators: *transposition* and *restricted-rebutting*. Transposition tackles the problems of inconsistency and incompleteness for grounded semantics, and the combination of both operators prevents these problems for preferred semantics.

Yet, due to questions over the reliability of the tool when enabling these operators, they were not used in this work and shall not be discussed further.

D.2.2 Architecture

One method of implementing argumentation splits the task into four distinct phases - the phases being related closely to the levels of Prakken's framework. Each phase is represented by a module in the ASPIC engine:

ArgCon creates individual arguments;

ArgVal implements the weakest/last link algorithm to assign a degree of strength to each argument;

ArgInt defines the notion of attack and defeat, and thus creates a network of interaction between the arguments;

ArgStat evaluates the status of each argument to determine which are justified.

In the above view of argumentation, the process starts by generating all possible/relevant arguments. These arguments are assessed individually to determine their degree of belief. A network (or graph) of argument relationships is produced, showing which arguments attack each other. Finally the network is used to generate a list of justified arguments. For example, for grounded semantics, this starts by including those arguments that are not attacked, and then recursively re-instating any defended arguments.

This view of argumentation is illustrated in Figure D.4 which is taken from an ASPIC project deliverable [154]. This diagram shows a potential relationship between the engine and other tools, which may be facilitated by the Argument Interchange Format (AIF).

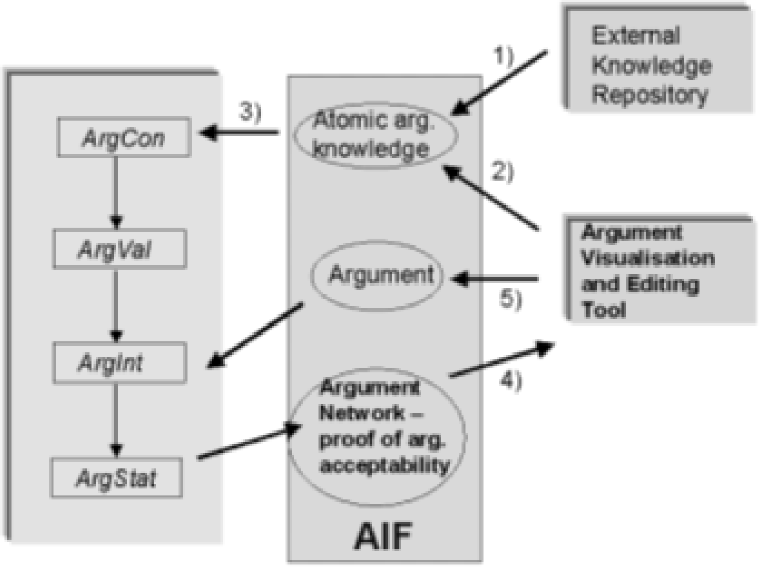

External Knowledge Repository
1)
3)
ArgCon
Atomic arg. knowledge
2)
ArgVal
Argument
5)
Argument Visualisation and Editing Tool
ArgInt
4)
Argument Network – proof of arg. acceptability
ArgStat
AIF

Figure D.4: Image taken from [154] demonstrating how the ASPIC argumentation engine might be used alongside other tools.

As discussed above, in Section D.1.5, the fourth process is modelled as a dialogue game. This is in keeping with the Loui's view of argumentation [39] that argumentation can be rational only if it is conducted inside a fair and effective dialectic disputation protocol.

Although Figure D.4 describes the knowledge store as separate, from a user's perspective it is not distinct from the engine.

D.2.3 Interfaces

The ASPIC argumentation engine has two user interfaces (UI), and one programmatic interface. Each will be discussed in turn, starting with the UIs.

Command line interface

The engine is written in JAVA. Therefore, it can be executed from the command line using JAVA. This interface allows the rules and facts to be saved in a text file, and then passed to the engine. Alternatively, the rules/facts can be written on the command line alongside the query. An example might be:

```
java -cp aspic-inference.0-4-10.bundled.jar org.aspic.inference.Engine
“a. b 0.8. ∼b.” “a.”
```

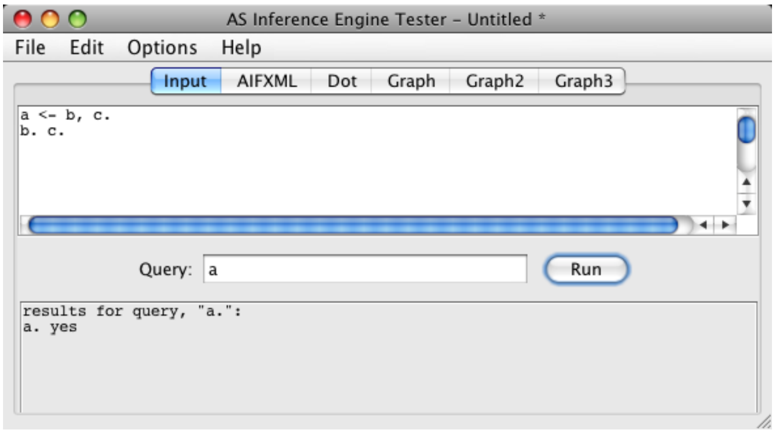
The output is uncomplicated. Each argument which produces a claim that matches the query is represented in the output. Alongside the argument is a simple ‘yes’ or ‘no’ to indicate which arguments are justified and which are false. For the above query (“a”) the result is simply:

```
a. yes
```

When using this interface, it is possible to change the semantics (grounded to preferred credulous), the valuation method (weakest link to last link), and switch on restricted rebutting and/or transposition. It is also possible to alter the form of the output, options include an XML form of the AIF, and a DOT representation that allows a graph of the justified argument(s) to be drawn.

Graphical interface

The same functionality provided by the command line interface, is likewise accessible via a simple tool with a graphical user interface (GUI). This time the information can be entered via the tool, see Figure D.5 for an example. Now it is possible to see the actual graphs, Figure D.6 provides an example. Other output representations, such as the AIF, remain displayed in textual form.



```
a <- b, c.
b. c.
```

Query: a

```
results for query, "a.":
a. yes
```

Figure D.5: ASPIC argumentation engine Graphical User Interface

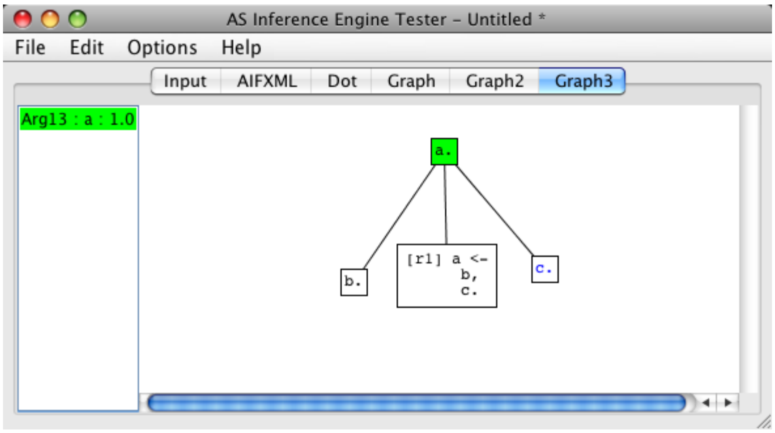


Figure D.6: One possible output from the ASPIC argumentation engine Graphical User Interface

JAVA Interface

Although the argumentation engine is not open source, the programmatic interface was made available. This is a series of classes which provide access to the same functionality as the graphical and command line interfaces. In addition, through this interface, it is possible to extend the engine by adding unique implementations of each of the classes. It therefore becomes possible to determine which set of arguments to display via a user interface, and in what manner to display them. An additional advantage of this interface, is that it allows the engine to be used as part of other tools and projects.

The interface provides a series of objects that model the basic elements of the PROLOG-like logic syntax, e.g. terms, and weak negation. Returning to the example from Section D.1.1 the PROLOG-like rule wet <- raining 0.7 can be written as:

```
new Rule(new Term("wet"), new ElementList(new Term("raining")), 0.7);
```

As the functionality is equivalent to that of the PROLOG-like syntax, it will not be discussed further.

D.3 Example of argumentation

In order to understand the behaviour of the tool, and the mechanism it uses to create then evaluate arguments a brief example is provided.

The example commences with the presumption there is a gene expression resource and a biologist. The basic intuition that "there is a reason to believe a gene is expressed in a structure when a gene expression resource suggests this is true" can be captured as:

```
expressed(G, S) <- resourceExpressed(G, S).
```

Thus, if the resource says the gene is not expressed:

```
~expressed(G, S) <- resourceNotExpressed(G, S).
```

The same is true for the biologist:

```
expressed(G,S) <- biologistExpressed(G, S).
~expressed(G, S) <- biologistNotExpressed(G, S).
```

Appropriate facts may include that the resource believes the gene *bmp4* is expressed in the structure called hindlimb, whereas the biologist disagrees:

```
resourceExpressed(bmp4, hindlimb).
biologistNotExpressed(bmp4, hindlimb).
```

If the query expressed(bmp4, hindlimb). is posed, the following arguments will be produced:

```
expressed(bmp4, hindlimb)
    resourceExpressed(bmp4, hindlimb).
    expressed(G, S) <- resourceExpressed(G, S).
```

and

```
∼expressed(bmp4, hindlimb)
        biologistNotExpressed(bmp4, hindlimb).
        ∼expressed(G, S) <- biologistNotExpressed(G, S).
```

The system will report that the query is justified. However, inverting the query to ask if the gene is not expressed also results in 'yes'. This result seems confusing, and yet the solution is simple. Both arguments are strict, and thus cannot be defeated. Consequently, both arguments are believed to be true, which is why both queries are answered positively.

If both arguments are assigned a degree of belief, this problem is resolved. If the facts are changed so that they are both assigned a belief of 0.8, the actual arguments inherit that score (assuming weakest link principle). At this point the status of the arguments depends on the semantics used. Grounded semantics are sceptical, and so both arguments will be defeated. Preferred credulous semantics are credulous, and so both arguments will be acceptable. Thus for grounded semantics, both queries will be answered negatively; however, they will both be answered positively for preferred credulous semantics.

If the arguments are made defeasible by assigning a degree of belief of 0.8 to the resource's fact and 0.7 to the biologist's, then the semantics are immaterial. The only justified argument suggests that the gene is expressed. Thus the first query results in a positive answer, and the second a negative response.

Undercut

If the above example is extended, with the information that the resource has had its security compromised we could have the following knowledge base (with the rule names made explicit):

```
[r1]expressed(G, S) <- resourceExpressed(G, S) 0.8.
[r2]~expressed(G, S) <- resourceNotExpressed(G, S) 0.8.
[r3]expressed(G, S) <- biologistExpressed(G, S) 0.8.
[r4]~expressed(G, S) <- biologistNotExpressed(G, S) 0.8.
resourceExpressed(bmp4, hindlimb) 0.8.
biologistNotExpressed(bmp4, hindlimb) 0.8.
[r5]~r1 <- resourceHacked.
[r6]~r2 <- resourceHacked.
resourceHacked 0.9.
```

Repeating the query, expressed(bmp4, hindlimb). results in the answer “no” as there is no justified argument supporting that conclusion. Although an argument for that conclusion may be generated, it will be undercut by an argument based on r5.

~expressed(bmp4, hindlimb). will receive a positive answer as an argument can be produced, and that argument must be justified as its only attacker is defeated.

D.3.1 Issues that may cause problems

All implementations of argumentation have flaws. This is due largely to the fact that there is no perfect theory which can be implemented. In this section, a few of the potential issues created by the selection of the ASPIC engine will be considered. A roundup of generic issues, such as floating arguments, can be found in [150, 38].

Caminada and Amgoud [156] highlight that many argumentation systems produce results that are neither complete nor consistent. ASPIC is not immune from this, unless the extra operators are enabled. As these operators occasionally caused the engine to fail, they were not enabled.

A more troublesome problem arises from the lack of aggregation. If an argument is stronger than another it wins any rebuttal. If there is one argument of strength 0.9 and an opposing argument of strength 0.8, the first argument will win perpetually. Yet, if there are four hundred opposing arguments, each with a strength of 0.8, then intuitively the sheer weight of numbers should cause the opposing arguments to win. Regrettably, this is not the case, as each argument is compared directly with the original argument, and thus is weaker. This has been considered and a solution implemented by others including Pollock [69]. However, no similar mechanism exists within the ASPIC engine.

A third complication is the implementation of weak defeat. \+(a) will be true when there is no fact or satisfied rule concluding a otherwise it will be false (∼a has no effect). In the following example assume that the contents of a resource are believed if they are not contradicted by a biologist:

```
expressed(G,T) <- resourceExpressed(G,T),
                  \+(biologistNotExpressed(G,T)).
biologistNotExpressed(bmp4, hindlimb) 0.8.
resourceExpressed(bmp4, hindlimb).
```

In this example, it should be clear that there is no argument supporting the conclusion the gene is expressed, because a biologist disagrees with the resource.

Now envisage an expert on gene expression being asked to review the opinion of the biologist. If (s)he disagrees, then the original biologist's opinion is no longer valid. The example is extended to include:

```
∼biologistNotExpressed(G, S) <- expertExpressed(G, S).
expertExpressed(bmp4, hindlimb).
```

In this situation, the only argument supporting the conclusion the biologist disagrees with the resource (biologistNotExpressed(bmp4, hindlimb)) is defeated. Thus it would seem reasonable to assume that as the resource is no longer contradicted, its opinion would be reinstated, i.e. expressed(bmp4, hindlimb) would be true. Yet this does not happen.

In reality biologistNotExpressed(bmp4, hindlimb) is defeated - it has to be as the expert's argument is strict. Yet \+(biologistNotExpressed(bmp4, hindlimb)) is not satisfied, and thus the resource cannot be trusted. The assumption is not satisfied because although biologistNotExpressed(bmp4, hindlimb) is defeated, there is still one argument supporting that conclusion and that is enough evidence for the argumentation engine to conclude that the assumption is invalid. This behaviour is not immediately intuitive.

Appendix E

Full list of comments made during the evaluation

Section 8.1.3 featured an abbreviated list of user comments obtained during the evaluation. The full list is below.

Presentation of arguments

- Illustration is extremely useful for an overview, although text is useful to determine strength of the argument developed;
- Keep first view simple and then have options to display more data if wanted;
- The diagrams and summaries seem intuitive and useful. Paragraphs pertaining to studies were hard to understand, but I'm a psychologist!
- Would like to have more explanation of argumentation process and why one argument wins over another;
- Should show arguments as an ontology: Not expressed or Expressed with parent and child terms (direct vs. inferred argument);
- Would like to be able to view each argument as a graph;
- A direct link to visual data would be helpful;
- Would like strengths added to graph of argument;

- Not knowing the biological processes the textual arguments are very hard to understand.

Issues of detail

- I may look at detailed descriptions if I need to do more work (dry or wet) and would like to be able to navigate different levels of detail where desired;

- The diagram of arguments is helpful but the detailed description is needed as well. The diagram is just the simplified version (summary) of the text and I don't trust the diagram alone;

- Would use personal expertise of system to re-query based on argument and look at more detail e.g. link to database would be useful;

- Would like to be able to view some sort of personal profile for each researcher to help user determine level of trust;

- Should show on the results page how any changes of trust in journals/databases/authors have altered the results and what changes in parameters have occurred.

The scenarios

- The Argument scenarios are organised in a logical way to 'vote' the final decision. I like it. One thing I can think of is the percent value allocated to the argument. Maybe they are too arbitrary/objective? Could users be given instructions on how to choose the confidence level?

- The task for the evaluation used genes that are well mapped - it would be problematic if there were a greater number of mappings to a gene;

- Not really comfortable with artificiality of example;.

Appendix F

Evaluation protocols

The protocols used for the evaluations discussed in Chapter 8 are provided here, they include:

- Script
- Time-error form
- Consent form
- Background questionnaire
- Argumentation scenarios and questionnaire
- Usability questionnaire

SEALIFE Evaluation Script

General

1. User reads and completes: consent form + background questionnaire
2. Mark forms with User Number
3. Make sure user happy with computer and controls
4. Make sure pop-up windows enabled on browser

Planning

1. Navigate to links page for user (go to TCM user interface)
2. User reads and begins scenario 1
3. Start timing and error observation
4. Record stage 5 timing
5. End timing
6. User begins scenario 2
7. Start timing and error observation
8. End timing
9. User answers Planning questionnaire

Argumentation

1. Navigate to Argumentation UI
2. User reads and begins scenario 1
3. Start timing and error observation
4. End timing
5. User begins scenario 2
6. Start timing and error observation
7. End timing
8. User answers Argumentation questionnaire

Questionnaire

1. User answers General questionnaire

SEALIFE Timing and Error Monitoring

Date: _ _ _ _ _ _

Start: _ _ _ _: _ _ _ _ : ._ _ _ 9.1 End: _ _ _.: _ _ _ _: _ _ _ _ 9.2 _ _ _ _

PLANNING

PART A: Manual Method: Errors:

10

1. Start step 1: _ _ _ _: ._ _ _ : _ _ _. 9.3 Step Error
2. Start of Step 5: _ _ _ _: ._ _ _ : _ _ _. 9.4
3. Blast 1 duration: _ _ _ _: ._ _ _ : _ _ _. 9.5
4. Blast 2 duration: _ _ _ _: ._ _ _ : _ _ _. 9.6
5. End time: _ _ _ _: ._ _ _ : _ _ _. 9.7

Part B – Automated Method

1. Start step 1: _ _ _ _: ._ _ _ : _ _ _. 9.8
2. End time: _ _ _ _: ._ _ _ : _ _ _. 9.9

ARGUMENTATION

Scenario 1

1. Start step 1: _ _ _ _: ._ _ _ : _ _ _.9.10
2. End time: _ _ _ _: ._ _ _ : _ _ _.9.11

Scenario 2

1. Start step 1: _ _ _ _: ._ _ _ : _ _ _.9.12
2. End time: _ _ _ _: ._ _ _ : _ _ _.9.13

Notes: 11

English is 1st Language: **Yes** □ **No** □ **Don't know** □ 9.14

SEALIFE Evaluation

Thank you for agreeing to participate in this evaluation.

We are evaluating two systems that are part of the SEALIFE Project.

1. The **Planning System** assists Biologists and Bioinformaticians to plan and execute bioinformatics workflows online. The system takes a bioinformatics resource and determines the available online processes that the resource can be submitted to.

2. The **Argumentation System** assists Biologists and Bioinformaticians in resolving conflicts between online gene expression databases on whether or not genes are expressed in particular tissues. The Edinburgh Mouse Atlas Gene Expression (EMAGE) database, Gene Expression Database (GXD) and Mouse Atlas of Gene Expression database (SAGE) are used for this evaluation.

This evaluation will ask you to work through scenarios using the systems and comment on the experience of using it. We will also ask your opinion on information presentation. There is no time element to the evaluation, and we will be on hand to answer questions and give any help if necessary.

No personally identifiable data will be stored, but we would be grateful if you would answer some questions on your professional background and experience.

- -

I agree to participate in the evaluation and to the collection and analysis of the data necessary for the evaluation.

Name: _ Signature: _

Date: _

SEALIFE Background Questionnaire ____

Area/s and level of training:

Please tick the boxes which reflect your qualification/training (more than one if appropriate):

Biology:	BSc. □		MSc. □		PhD. □		Other ________
Bioinformatics:	BSc. □		MSc. □		PhD. □		Other ________
Genetics:	BSc. □		MSc. □		PhD. □		Other ________
Computing/IT:	BSc. □		MSc. □		PhD. □		Other ________
Medicine:	MB.ChB. □		MSc. □		MD./PhD. □		Other ________
Other: (Please specify)	______ □		____ □		____ □		____________

Current and Previous Posts:

Please indicate in years the approximate time spent in post (indicate current post with asterisk):

	Biology	Bioinformatics	Computing/IT:	Other: _________ (Please specify)
Undergrad:	□	□	□	□
Postgrad/Research:	□	□	□	□
Industry:	□	□	□	□

Tools and Databases

Please indicate your levels of familiarity with the following tools and databases:

XSPAN

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.1

EMAGE

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.2

GXD

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.3

Blast

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.4

UniProt

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.5

CGAP

Never Used **Very Familiar**

1 2 3 4 5 6 7 8 9 N/A 1.6

Molecular Biology and Bioinformatics Background

How knowledgeable are you about Gene Expression?

No knowledge								Expert		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	1.7

How knowledgeable are you about **ontologies** ?

No knowledge								Expert		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	1.8

Journals

Please tick to indicate your reading of the following journals:

Journal	**Read all of journal**	**Read some papers**	**Know journal but don't read it**	**Never heard of journal**
Development	□	□	□	□
Mechanisms of Development	□	□	□	□
Science	□	□	□	□
Biochimica et Biophys. Acta	□	□	□	□
Nature	□	□	□	□
EMBO	□	□	□	□
Developmental Biology	□	□	□	□
Gene Expression Patterns	□	□	□	□
Molecular Biology of the Cell	□	□	□	□

SEALIFE Argumentation Evaluation Scenario

Argumentation

Biological resources publish experimental results and for various reasons, these experiments often give conflicting conclusions. Argumentation is a mechanism to help resolve such conflict. In this evaluation we shall use 3 resources for in-situ gene expression in the developmental mouse. Each resource will provide all the experimental data it contains for a particular gene-structure pair. This data will be used to create arguments, weighted according to trust factors determined by you, for and against the gene being expressed in the structure. The arguments will be presented to you after the argumentation process has terminated.

Evaluation

For the evaluation, you will be asked to work through some arguments using a prototype of the argumentation user interface, and then answer some questions and comment on the experience.

Please feel free to ask questions or comment at any stage of the evaluation.

Section 1

The first screen comprises the interface with some pre-loaded gene-structure combinations that will be used for the argumentation process.

1. On the start page select the **Telencephalon TS15 EMAP:1212 Bmp4** item from the cart on the right of the screen using the appropriate **Select** button.

2. For this scenario, levels of trust of journals that publish gene expression papers and the gene expression databases are left at default values;

3. Click the **Next Page** button;

4. You are presented with references related to the gene expression data;
   For this scenario, leave levels indicating your trust of the researchers/laboratories at default values;
   You can click on the PubMed button to see journal abstracts where available (close the PubMed windows/tabs to continue);

5. Click the **Next Page** button;

6. The final page shows the conclusion of the arguments, using a schematic of the argument and natural language descriptions of each indicator;

7. Which do you think is the strongest argument/s? _

Do you think the Bmp4 gene is expressed in EMAPA:1212 ? **Yes** □ 8.1
No □
Don't know □

What is your decision based on? _

_ _

_ _

_ 8.2

8. Close the second and third pages to return to the start page;

Section 2

This scenario follows a similar path, but illustrates the effect on the argument of not using one of the databases and choosing low vs. high levels of confidence the researchers.

1. On the start page select the **Telencephalon TS15 EMAP:1212 Bmp4** item from the cart on the right of the screen using the appropriate **Select** button.

2. De-select **CGAP** from the databases list;

3. For the **Development** journal - select a level of trust of **25%**

4. Click the **Next Page** button.

5. In the list of **Experiments** select a 25% level of trust for the **Ishibashi & McMahon, 2002 [PMID:12361972];**

6. Click the **Next Page** button.

7. The final page shows the conclusion of the arguments, using a schematic of the argument and natural language descriptions of each indicator.

8. Which do you think is the strongest argument/s? _

Do you think the Bmp4 gene is expressed in EMAPA:1212 ? **Yes** □ 8.3

No □

Don't know □

What is your decision based on? _

_ _

_ _

_ 8.4

9. Close the second and third pages to return to the start page;

Please answer the following questions based on your use of the system:

Were the questions on trust of journals and researchers asked of you

Too few **Too many**

1 2 3 4 5 6 7 8 9 N/A 8.5

Would the option to hide the questions on trust of journals and authors be

Undesirable **Very desirable**

1 2 3 4 5 6 7 8 9 N/A 8.6

Was the amount of information presented in the arguments

Too little **Too much**

1 2 3 4 5 6 7 8 9 N/A 8.7

How well did you understand the information presented in the arguments

Not at all **Completely**

1 2 3 4 5 6 7 8 9 N/A 8.8

Did you find the diagrams of the arguments helpful

Totally unhelpful **Very helpful**

1 2 3 4 5 6 7 8 9 N/A 8.9

Can you suggest any other resources that could be included in the system?

_ _

_ 8.10

Section 3

Journal Trust

Please indicate your confidence in the following journals by ranking them from **1** (highest confidence) to **9** (lowest confidence) in the appropriate box.

Journals	Rank (1 - 9)	Don't know
Development	□	□
Mechanisms of Development	□	□
Science	□	□
Biochimica et Biophysica Acta	□	□
Nature	□	□
EMBO	□	□
Developmental Biology	□	□
Gene Expression Patterns	□	□
Molecular Biology of the Cell	□	□

8.11 - 8.19

The following list of authors publish in the field of in-situ gene expression for the developmental mouse. Please put a tick in the box next to any authors you recognise.

Authors	Recognise
Furuta Y	□
Hogan BL	□
Trainor PA	□
Hebert JM	□
Martin GR	□
Niswander L	□

8.20 - 8.25

Section 4

The following is a statement generated by the argumentation system. Please read it and answer the questions below.

"There is experimental evidence the gene is not expressed. There is no reason to doubt the annotation: (Negative expression can be propagated down). TELENCEPHALON is part of FUTURE BRAIN. EMAGE:998 has a textual annotation showing Bmp4 is not expressed in FUTURE BRAIN TS15 (EMAP:1199). STRENGTH=80%."

What can you infer from the above argument? _

_ _

_ _

_ 8.26

How easy is this argument to understand?

Very confusing						**Very straightforward**				
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	8.27

How intuitive do you find the inference drawn in the argument?

Not intuitive							**Very intuitive**			
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	8.28

How could this method of presentation be improved? _

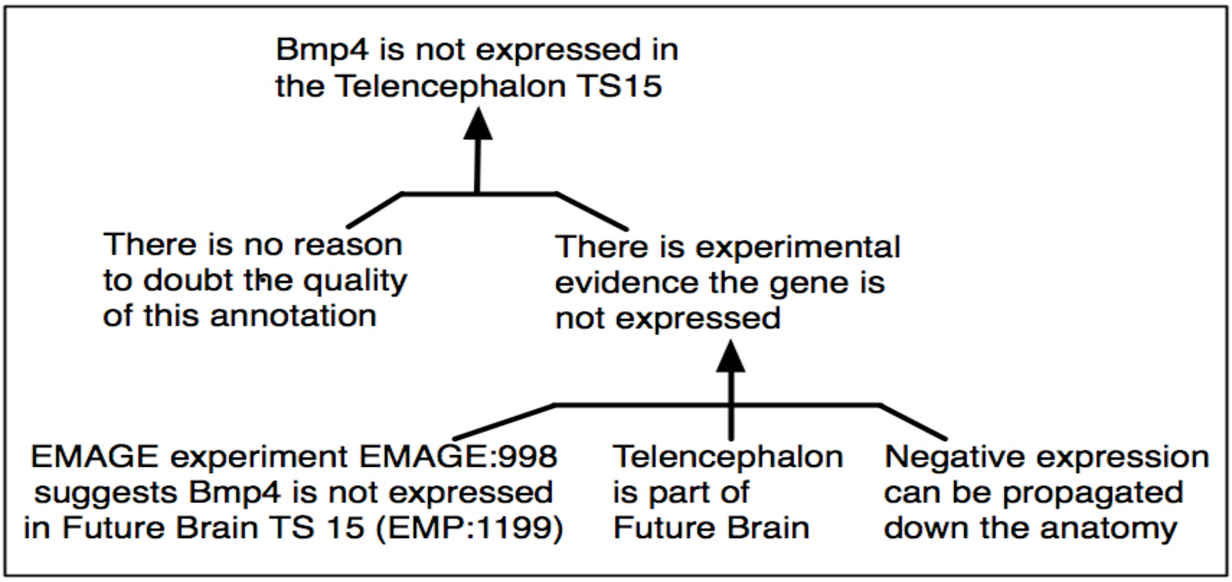
_ _

_ _

_ 8.29

Section 5

Please look at the argument contained within this graph and answer the questions below.


Bmp4 is not expressed in the Telencephalon TS15
There is no reason to doubt the quality of this annotation
There is experimental evidence the gene is not expressed
EMAGE experiment EMAGE:998 suggests Bmp4 is not expressed in Future Brain TS 15 (EMP:1199)
Telencephalon is part of Future Brain
Negative expression can be propagated down the anatomy

What can you infer from the above graph? ___________________________________

____

____ 8.30

How easy is this argument to understand?

Very confusing						Very straightforward				8.31
1	2	3	4	5	6	7	8	9	N/A	

How intuitive do you find the inference drawn in the argument?

Not intuitive						Very intuitive				8.32
1	2	3	4	5	6	7	8	9	N/A	

How could this method of presentation be improved? ______________________________

____

____ 8.33

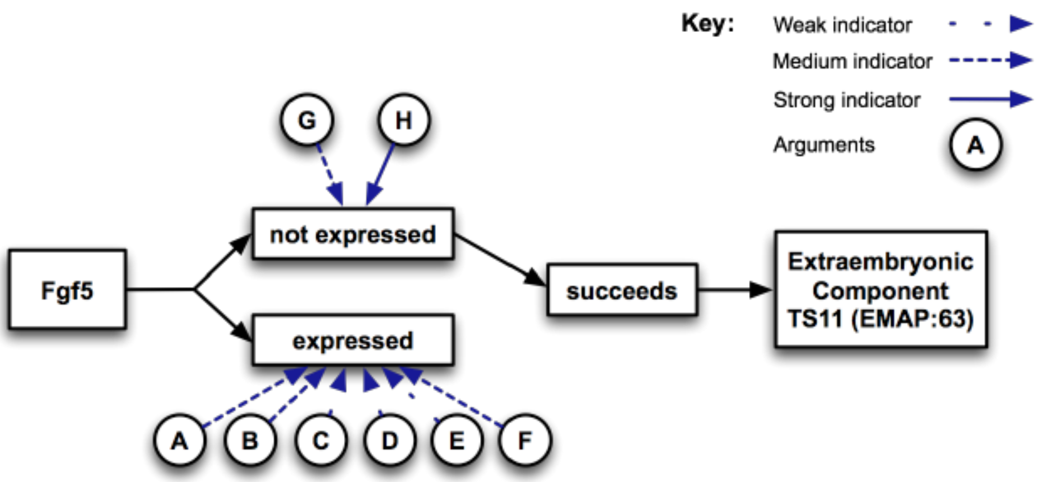
Which form of presentation do you prefer? **Natural language paragraph** □ 8.34

(tick one) **Graph** □

Don't know □

Section 6

Please look at the following result of an argumentation process and answer the questions below.



What does this diagram suggest to you ?	**Fgf5 is expressed**	□
(tick one)	**Fgf5 is not expressed**	□
	Don't know	□

8.35

Looking at the diagram above, what do you think are the strengths of the following arguments?

	Strong	**Medium**	**Weak**	**Don't know**	
Argument **H**	□	□	□	□	8.36
Argument **C**	□	□	□	□	8.37
Argument **A**	□	□	□	□	8.38

How easy is this diagram to understand?

Very easy							**Very hard**			
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	8.39

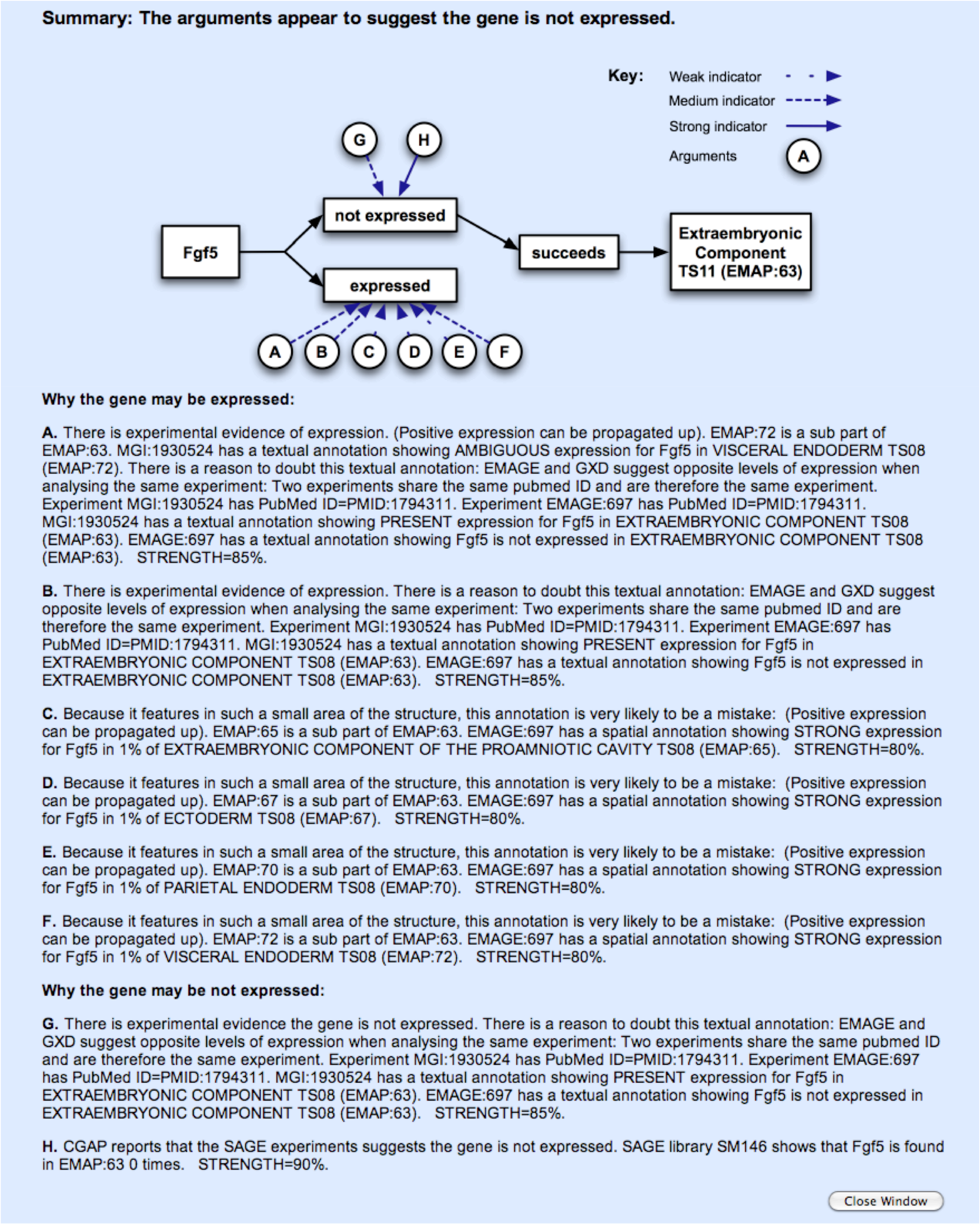
How could this method of presentation be improved? _

_ _

_ _

_ 8.40

Section 7



Summary: The arguments appear to suggest the gene is not expressed.

Why the gene may be expressed:

A. There is experimental evidence of expression. (Positive expression can be propagated up). EMAP:72 is a sub part of EMAP:63. MGI:1930524 has a textual annotation showing AMBIGUOUS expression for Fgf5 in VISCERAL ENDODERM TS08 (EMAP:72). There is a reason to doubt this textual annotation: EMAGE and GXD suggest opposite levels of expression when analysing the same experiment: Two experiments share the same pubmed ID and are therefore the same experiment. Experiment MGI:1930524 has PubMed ID=PMID:1794311. Experiment EMAGE:697 has PubMed ID=PMID:1794311. MGI:1930524 has a textual annotation showing PRESENT expression for Fgf5 in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). EMAGE:697 has a textual annotation showing Fgf5 is not expressed in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). STRENGTH=85%.

B. There is experimental evidence of expression. There is a reason to doubt this textual annotation: EMAGE and GXD suggest opposite levels of expression when analysing the same experiment: Two experiments share the same pubmed ID and are therefore the same experiment. Experiment MGI:1930524 has PubMed ID=PMID:1794311. Experiment EMAGE:697 has PubMed ID=PMID:1794311. MGI:1930524 has a textual annotation showing PRESENT expression for Fgf5 in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). EMAGE:697 has a textual annotation showing Fgf5 is not expressed in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). STRENGTH=85%.

C. Because it features in such a small area of the structure, this annotation is very likely to be a mistake: (Positive expression can be propagated up). EMAP:65 is a sub part of EMAP:63. EMAGE:697 has a spatial annotation showing STRONG expression for Fgf5 in 1% of EXTRAEMBRYONIC COMPONENT OF THE PROAMNIOTIC CAVITY TS08 (EMAP:65). STRENGTH=80%.

D. Because it features in such a small area of the structure, this annotation is very likely to be a mistake: (Positive expression can be propagated up). EMAP:67 is a sub part of EMAP:63. EMAGE:697 has a spatial annotation showing STRONG expression for Fgf5 in 1% of ECTODERM TS08 (EMAP:67). STRENGTH=80%.

E. Because it features in such a small area of the structure, this annotation is very likely to be a mistake: (Positive expression can be propagated up). EMAP:70 is a sub part of EMAP:63. EMAGE:697 has a spatial annotation showing STRONG expression for Fgf5 in 1% of PARIETAL ENDODERM TS08 (EMAP:70). STRENGTH=80%.

F. Because it features in such a small area of the structure, this annotation is very likely to be a mistake: (Positive expression can be propagated up). EMAP:72 is a sub part of EMAP:63. EMAGE:697 has a spatial annotation showing STRONG expression for Fgf5 in 1% of VISCERAL ENDODERM TS08 (EMAP:72). STRENGTH=80%.

Why the gene may be not expressed:

G. There is experimental evidence the gene is not expressed. There is a reason to doubt this textual annotation: EMAGE and GXD suggest opposite levels of expression when analysing the same experiment: Two experiments share the same pubmed ID and are therefore the same experiment. Experiment MGI:1930524 has PubMed ID=PMID:1794311. Experiment EMAGE:697 has PubMed ID=PMID:1794311. MGI:1930524 has a textual annotation showing PRESENT expression for Fgf5 in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). EMAGE:697 has a textual annotation showing Fgf5 is not expressed in EXTRAEMBRYONIC COMPONENT TS08 (EMAP:63). STRENGTH=85%.

H. CGAP reports that the SAGE experiments suggests the gene is not expressed. SAGE library SM146 shows that Fgf5 is found in EMAP:63 0 times. STRENGTH=90%.

Presented with the above results from the argumentation system, which sections would you use to make your decision on whether the gene was expressed or not?

(please tick as many as needed) **Use**

The summary at the top ☐ 8.41

The diagram of the arguments ☐ 8.42

The detailed descriptions of the arguments ☐ 8.43

Section 8

Comments on the Argumentation Scenarios

Please write any comments you have in the space below.

– –

– –

– –

– –

– –

– –

– –

– –

– –

– 8.44

SEALIFE Usability Evaluation Questionnaire _ _ _ _

This section comprises questions on the overall system, both the Task Composition and Argumentation sections. Please circle the numbers which most appropriately reflect your impressions about using the SEALIFE system. Not Applicable = N/A. Please feel free to write any comments you have on the form.

Reactions to System Overall

What is your overall reaction to the system

Terrible **Wonderful**

1 2 3 4 5 6 7 8 9 N/A 2.1

Frustrating **Satisfying**

1 2 3 4 5 6 7 8 9 N/A 2.2

Dull **Stimulating**

1 2 3 4 5 6 7 8 9 N/A 2.3

Difficult **Easy**

1 2 3 4 5 6 7 8 9 N/A 2.4

Inadequate power **Adequate power**

1 2 3 4 5 6 7 8 9 N/A 2.5

Rigid **Flexible**

1 2 3 4 5 6 7 8 9 N/A 2.6

Screen

Were the screen layouts helpful?

Not at all **Very much**

1 2 3 4 5 6 7 8 9 N/A 3.1

Sequence of screens

Confusing **Clear**

1 2 3 4 5 6 7 8 9 N/A 3.2

Terminology and System Information.

Use of terms throughout system

Inconsistent **Consistent**

1 2 3 4 5 6 7 8 9 N/A 4.1

Does the terminology relate well to the work you are doing?

Unrelated **Well related**

1 2 3 4 5 6 7 8 9 N/A 4.2

Messages which appear on screen

Inconsistent **Consistent**

1 2 3 4 5 6 7 8 9 N/A 4.3

Messages which appear on screen

Confusing **Clear**

1 2 3 4 5 6 7 8 9 N/A 4.4

Does the computer keep you informed about what it is doing?

Never **Always**

1 2 3 4 5 6 7 8 9 N/A 4.5

Error messages

Unhelpful **Helpful**

1 2 3 4 5 6 7 8 9 N/A 4.6

System Capabilities

System speed

Too slow								**Fast enough**		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	5.1

Correcting your mistakes

Difficult								**Easy**		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	5.2

Are the needs of both experienced and inexperienced users taken into account?

Never								**Always**		
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**N/A**	5.3

Comments on overall system

Please write any comments you have in the space below. 5.4

Appendix G

Argudas evaluation protocols

The protocol used for the Argudas evaluation is provided here, including:

- Consent form
- Scenario
- Usability questionnaire
- Time-error form

Argudas Evaluation

What is Argudas?
Argudas is a system designed to help a user explore the gene expression information contained within the EMAGE and GXD databases. Additionally, for the adult mouse, data from GENSAT and the Allen Brain Atlas is integrated.

What is the purpose of this exercise?
This evaluation will explore how easy Argudas is to use.

What instructions will I be given?
Step-by-step instructions to walk you through Argudas appear on the next page. Please follow them carefully. If at any stage you are unsure of what to do, please ask. Likewise, please make any comments you wish about Argudas or this exercise.

If you are happy to progress with the exercise please sign and date below:

Signature ____ Date ______________

You may withdraw from this exercise at any time you wish.

Please follow the steps described below:

1. If the browser is not already there, navigate to: http://lxbisel.macs.hw.ac.uk:8080/Argudas

2. You will try to discover if *bmp4* is expressed in the future brain in TS13:
    a. Click 'Show / hide help'
    b. If you wish more information click 'For more instructions and help click here'
    c. Enter the appropriate query and click 'submit'

3. Do you think there is sufficient information in the results to determine if *bmp4* is expressed?

    Yes / No / Don't know

    Why do you think this?

    ____

    ____

    ____

4. Ask Argudas to provide more information

5. Once you see the two new tables:
    a. If you wish assistance understanding the output, click 'For more help interpreting the tables click here'.
    b. Otherwise, do you think *bmp4* is expressed?

        Yes / No / Don't know

    c. What is your decision based on?

        ____

        ____

        ____

6. Please now fill in the questionnaire.

Please circle your answer to the following questions:

1. Please rate Argudas’ response time:

Excellent				*Bad*	*Don’t know*
1	2	3	4	5	0

2. Please rate Argudas’ appearance:

Excellent				*Bad*	*Don’t know*
1	2	3	4	5	0

3. Do you think the extra information provided by clicking `show / hide help’ is necessary? *Yes* / *No*

4. Did you find the first results table easy to understand?

Very easy				*Very difficult*	*Don’t know*
1	2	3	4	5	0

5. Did you find the second results table easy to understand?

Very easy				*Very difficult*	*Don’t know*
1	2	3	4	5	0

6. Is there any extra information you would like to add to the second table?

7. Did you find the third results table easy to understand?

Very easy				*Very difficult*	*Don’t know*
1	2	3	4	5	0

8. Is there any extra information you would like to add to the third table?

9. Why did you use the Argudas help page?

10. Please rate the usefulness of the help page?

Very helpful				*Not helpful*	*Don’t know*
1	2	3	4	5	0

11. What is missing from the help page?

12. Any comments you wish to make about Argudas?

The questionnaire is now complete: THANKS FOR YOUR ASSISTANCE!

Protocol

User no: ____ Date: ___ / ___ / 2011 1st language English: Yes / No

###

Start time: ____________

Did they view help page after switching on help ‘boxes’: Yes / No

Did they enter a ‘correct’ query: Yes / No

Bad query = ____

Did they view help page when presented with tables 2 & 3: Yes / No

End time: ____________ Time taken: ____________

###

Comments on their behaviour:

###

Comments they made:

References

[1] T.J.M. Bench-Capon and P. E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10–15):619–641, July–October 2007.

[2] D. Walton, C. Reed, and F. Macagno. *Argumentation schemes.* Cambridge University Press, New York, NY, USA, 2008.

[3] B. Garssen. *Crucial concepts in argumentation theory*, chapter Argumentation schemes, pages 81–99. Amsterdam University Press, 2001.

[4] *Oxford dictionary of English*, chapter Definition of 'defeasible'. Oxford University Press, 3rd edition edition, 2010.

[5] T.J.M. Bench-Capon and H. Prakken. *Information Technology & Lawyers: Advanced technology in the legal domain, from challenges to daily routine*, chapter Argumentation, pages 61–80. Springer Netherlands, 2006.

[6] C. Trojahn, P. Quaresma, and R. Vieira. *Law, ontologies and the semantic web - channelling the legal information flood*, volume 188 of *Frontiers in artificial intelligence and applications*, chapter Matching law ontologies using an extended argumentation framework based on confidence degrees, pages 133 – 144. IOS Press, 2009.

[7] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, and V. Patkar. Argumentation-based inference and decision making - a medical perspective. *IEEE Intelligent Systems*, 22(6):34–41, November/December 2007.

[8] I. Rahwan, S.D. Ramchurn, N.R. Jennings, P. M^{c}Burney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *The Knowledge Engineering Review*, 18(4):343–375, 2003.

[9] B.R. Jefferys, L.A. Kelly, M.J. Sergot, J. Fox, and M.J.E. Sternberg. Capturing expert knowledge with argumentation: a case study in bioinformatics. *Bioinformatics*, 22(8):924–933, 2006.

[10] A. Adúriz-Bravo, L. Bonan, L.G. Galli, A.R. Chion, and E. Meinardi. Scientific argumentation in pre-service biology teacher education. *Eurasia journal of mathematics, science and technology education*, 1(1):76–83, November 2005.

[11] F. Grasso, A. Cawsey, and R. Jones. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, 53(6):1077–1115, December 2000.

[12] N. Green. Representing normative arguments in genetic counseling. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 64–68, 2006.

[13] trans. E.S. Forster Aristotle. *Topics.* Loeb Classical Library. Harvard University Press, Cambridge, Mass. USA, 1960.

[14] D. Walton. *Fundamentals of critical argumentation.* Critical reasoning and argumentation. Cambridge University Press, New York, NY, USA, 2006.

[15] B.E. Gronbeck. *Readings in argumentation*, chapter From argument to argumentation: fifteen years of identity crisis, pages 17–32. Studies of argumentation in pragmatics and discourse analysis. Walter de Gruyter & Co, Berlin, 1992.

[16] P. Besnard and A. Hunter. *Elements of argumentation.* The MIT Press, 2008.

[17] C. Perelman and L. Olbrechts-Tyteca. *The new rhetoric: a treatise on argumentation.* University of Notre Dame Press, 1969.

[18] H. Prakken, C. Reed, and D. N. Walton. Argumentation schemes and burden of proof. In F. Grasso, C. Reed, and G. Carenini, editors, *Workshop notes of the 4th International Workshop on Computational Models of Natural Argument (CMNA2004)*, pages 81–86, Valencia, Spain, August 2004.

[19] D.N. Walton and E.C.W. Krabbe. *Commitment in dialogue: basic concepts of interpersonal reasoning.* SUNY series in logic and language. SUNY, 1995.

[20] D. Walton and D.M. Godden. Informal logic and the dialectical approach to argument. In H.V. Hansen and R.C. Pinto, editors, *Reason reclaimed*, pages 3–17. Newport News, Virginia, VA: Vale Press, 2007.

[21] D. Walton. Justification of argumentation schemes. *Australasian Journal of Logic*, 3:1–13, July 2005.

[22] D.J. O'Keefe. Two concepts of argument. *The Journal of the American Forensic Association*, 13(3):121–128, 1977.

[23] trans. T M^{c}Carthy J. Habermas. *The theory of communicative action: reason and the rationalization of society*, volume 2. Beacon Press, 1985.

[24] F.H. van Eemeren and R. Grootendorst. *Argumentation, communication and fallacies: a pragma-dialectical perspective.* Routledge, 1992.

[25] C. Reed and F. Grasso. Recent advances in computational models of natural argument. *International Journal of Intelligent Systems*, 22(1):1–15, January 2007.

[26] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, April 1980.

[27] D. M^{c}Dermott and J. Doyle. Non-monotonic logic 1. *Artificial Intelligence*, 13:41–72, April 1980.

[28] K.L. Clark. *Logics and Data Bases*, chapter Negation as failure, pages 293–322. Plenum Press, New York, 1978.

[29] John M^{c}Carthy. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2):27–39, April 1980.

[30] F. Lin and Y. Shoham. Argument systems: a uniform basis for nonmonotonic reasoning. In *First International Conference on Knowledge Representation and Reasoning (KR '89)*, pages 224–255. Morgan Kaufmann, 1989.

[31] P.M. 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, September 1995.

[32] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1-2):63–101, June 1997.

[33] C. Reed, D. Walton, and F. Macagno. Argument diagramming in logic, law and artificial intelligence. *Knowledge Engineering Review*, 22(1):87–109, March 2007.

[34] A. de Moor and M. Aakhus. Argumentation support: from technologies to tools. *Communications of the ACM*, 49(3):93–98, March 2006.

[35] D.W. Glasspool, J. Fox, F.D. Castillo, and V.E.L. Monagham. Interactive decision support for medical planning. In M. Dojat, E. Keravnou, and P. Barahona, editors, *Artificial intelligence in medicine: 9th conference on Artificial Intelligence in Medicine in Europe (AIME 2003)*, Lecture Notes in Artificial Intelligence (LNAI 2780), pages 335–339, Berlin, 2003. Springer-Verlag.

[36] D. Walton and D. M. Godden. *Considering Pragma-Dialectics*, chapter The impact of argumentation on artificial intelligence, pages 287–299. Lawrence Erlbaum Associates, 2006.

[37] T.F. Gordon. Hybrid reasoning with argumentation schemes. In *Proceedings of the 8th Workshop on Compuational Models of Natural Argument (CMNA 08)*, pages 16–25, July 2008.

[38] H. Prakken and G. Vreeswijk. *Handbook of philosophical logic, second edition*, volume 4, chapter Logics for defeasible argumentation, pages 219–318. Kluwer Academic Publishers, 2002.

[39] R.P. Loui. Process and policy: resource-bounded non-demonstrative reasoning. *Computational Intelligence*, 14(1):1–38, Februrary 1998.

[40] B. Verheij. Deflog - a logic of dialectical justification and defeat. Technical report, Department of Metajuridica, Universiteit Maastricht, 2000.

[41] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.

[42] H. Prakken. From logic to dialectics in legal argument. In *Proceedings of the 5th International Conference on Artificial Intelligence and Law (ICAIL '95)*, pages 165–174, New York, NY, USA, 1995. ACM Press.

[43] T.F. Gordon, H. Prakken, and D. Walton. The carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10–15):875–896, July–October 2007.

[44] S. Shapiro. Classical logic. In E.N. Zalta, editor, *The Stanford encyclopedia of philosophy.* Fall 2008.

[45] J.L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57(1):1–42, September 1992.

[46] M.A. Gilbert. *Coalescent argumentation.* Routledge, 1997.

[47] C. Reed and D. Walton. Towards a formal and implemented model of argumentation schemes in agent communication. In I. Rahwan, P. Moraitis, and C. Reed, editors, *Autonomous Agents and Multi-agent Systems*, volume 11, pages 173–188. Kluwer Academic Publishers, September 2005.

[48] B. Verheij. Logic, context and valid inference. or: can there be a logic of law? In H.J. van den Herik, M.F. Moens, J. Bing, B. van Buggenhout, J. Zeleznikow, and C. A. F. M. Grütters, editors, *Legal Knowledge Based Systems. JURIX 1999: The twelfth conference*, pages 109–121, Nijmegen: Gerard Noodt Institute, 1999.

[49] G.F. Luger and W.A. Stubblefield. *Artificial Intelligence: structures and strategies for complex problem solving, 3rd edition*, pages 93–96. Addison-Wesley, 1998.

[50] D.N. Walton. Abductive, presumptive and plausible arguments. *Informal Logic*, 21(2):141–169, 2001.

[51] H.L.A. Hart. The ascription of responsibility and rights. In *Proceedings of the Aristotelian Society*, volume 49, pages 171–194, 1948.

[52] R. Koons. Defeasible reasoning. *The Stanford encyclopedia of philosophy (fall 2008 edition)*, 2008.

[53] J.L. Pollock. *Reasoning: studies of human inference and its foundations*, chapter Defeasible reasoning, pages 451–470. Cambridge University Press, 2008.

[54] G. Booch. *Object-oriented analysis and design with applications, 2nd edition.* Benjamin/Cummings, 1994.

[55] D.N. Walton. *Argumentation schemes for presumptive reasoning (Studies in argumentation series).* Lawrence Erlbaum Associates, Mawah, NJ, USA, 1996.

[56] J.A. Blair. Walton's argumentation schemes for presumptive reasoning: a critique and development. *Argumentation*, 15(4):365–379, November 2001.

[57] M. Kienpointner. *Alltagslogik. Struktur und funktion von argumentationsmustern.* Stuttgart-Bad Cannstatt: Frommann-Holzboog, 1992.

[58] A.C. Hastings. *A reformulation of the modes of reasoning in argumentation.* PhD thesis, Northwestern University, Evanston, IL, USA, 1963.

[59] D.M. Godden and D. Walton. Advances in the theory of argumentation schemes and critical questions. *Informal Logic*, 27(3):267–292, 2007.

[60] D.N. Walton and C.A. Reed. Argumentation schemes and defeasible inferences. In G. Carenini, F. Grasso, and C. Reed, editors, *Working notes of the ECAI 2002 Workshop on Computational Model of Natural Argument*, pages 45–55, 2002.

[61] D. Hitchcock. Does the traditional treatment of enthymemes rest on a mistake? *Argumentation*, 12(1):15–37, 1998.

[62] B. Verheij. Dialectical argumentation with argumentation schemes: an approach to legal logic. *Artificial Intelligence and Law*, 11(2–3):167–195, January 2003.

[63] F. Bex, H. Prakken, C. Reed, and D. Walton. Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. *Artificial Intelligence and Law*, 11(2–3):125–165, January 2003.

[64] C. Chesñevar, J. M^{c}Guinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott. Towards an argument interchange format. *Knowledge Engineering Review*, 21(4):293–316, December 2006.

[65] I. Rahwan, C. Reed, and F. Zablith. On building argumentation schemes using the argument interchange format. In *Working notes of the 7th Workshop on Computational Models of Natural Argument (CMNA 2007)*, pages 49–56, Hyderabad, 2007.

[66] C. Reed and G. Rowe. Araucaria: software for puzzles in argument diagramming and XML. Technical report, Department of Applied Computing, University of Dundee, 2001.

[67] D. Walton and T.F. Gordon. Computational models of legal argument. In P.E. Dunne and T. Bench-Capon, editors, *International workshop on argumentation in artificial intelligence and law*, IAAIL Workshop series, pages 103–111, Nijmegen, The Netherlands, 2005. Wolf Legal Publishers.

[68] J.L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, October–December 1987.

[69] J.L. Pollock. Defeasible reasoning with variable degrees of justification 2. http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Degrees.pdf, June 2002.

[70] S.E. Toulmin. *The uses of argument.* Cambridge University Press, 1958.

[71] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 7:25–75, 1997.

[72] L. Amgoud, C. Cayrol, M.C. Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093, October 2008.

[73] L. Amgoud and H. Prade. Using arguments for making decisions: a possibilistic logic approach. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 10–17, Arlington, Virginia, United States, 2004. AUAI Press.

[74] P. Krause, S. Ambler, M. Elvang-Goransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11(1):113–131, February 1995.

[75] L. Amgoud amd C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, March 2002.

[76] G.R. Simari and R.P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53(2–3):125–157, Februrary 1992.

[77] G. A. W. Vreeswijk. The feasibility of defeat in defeasible reasoning. In J.F. Allen, R. Fikes, and E. Sandwell, editors, *Proceedings of the second international conference on principles of knowledge representation and reasoning*, pages 526–534. Morgan Kaufmann, April 1991.

[78] T.J.M. Bench-Capon and P.E. Dunne. Value based argumentation frameworks. Technical Report ULCS-02-001, University of Liverpool, 2002.

[79] H.V. Hansen. The straw thing of fallacy theory: the standard definition of 'fallacy'. *Argumentation*, 16(2):133–155, June 2002.

[80] R.H. Johnson and J.A. Blair. *Logical self-defense (3rd edition).* McGraw-Hill, 1994.

[81] M. Williams and J. Williamson. Combining argumentation and bayesian nets for breast cancer prognosis. *Journal of Logic, Language and Information*, 15(1–2):155–178, July 2006.

[82] Y. Dimopoulos and A. Torres. Graph theoretical structures in logic programs and default theories. *Theoretical Computer Science*, 170(1–2):209–244, December 1996.

[83] J.H. Wigmore. *The principles of judicial proof: or, the process of proof as given by logic, psychology, and general experience and illustrated in judicial trials.* Little, Brown, 2nd edition, 1931.

[84] G. Rowe, F. Macagno, C. Reed, and D. Walton. Araucaria as a tool for diagramming arguments in teaching and studying philosophy. *Teaching Philosophy*, 29(2):111–124, 2006.

[85] T.J. van Gelder. Argument mapping with Reason!Able. In *The American Philosophical Association Newsletter on Philosopohy and Computers*, pages 85–90, 2002.

[86] D. Suthers, A. Weiner, J. Connelly, and M. Paolucci. Belvedere: engaging students in critical discussion of science and public policy issues. In *AI-ED 95, the 7th World Conference on Artificial Intelligence in Education*, pages 266–273, August 1995.

[87] S. Adams. Investigation of "convince me" computer environment as a tool for critical thinking and public policy issues. *Journal of Interactive Learning Research*, 14(3):263–283, July 2003.

[88] E.J. Conklin. The IBIS manual a short course in IBIS methodology.

[89] T. Rodden. A survey of CSCW systems. *Interacting with Computers*, 3(3):319–353, December 1991.

[90] E.J. Conklin and W. Weil. Wicked problems: naming the pain in organizations.

[91] J. Conklin and M.L. Begeman. gIBIS: a hypertext tool for exploratory policy discussion. In *Proceedings of the 1998 ACM conference on computer-supported cooperative work*, CSCW '88, pages 140–152, New York, NY, USA, 1998. ACM.

[92] J. Conklin. *Visualizing argumentation: software tools for collaborative and educational sense-making*, chapter Dialog mapping: reflections on an industrial strength case study. Springer Verlag: London, 2003.

[93] S. Buckingham Shum. There's nothing like a good argument. *IEEE Software*, 24(5):21–23, September/October 2007.

[94] B. Verheij. Automated argument assistance for lawyers. In *Proceedings of the 7th international conference on artificial intelligence and law*, ICAIL '99, pages 43–52, New York, NY, USA, 1999. ACM.

[95] T.F. Gordon. Visualizing Carneades argument graphs. *Law, probability & risk*, 6(1-4):109–117, October 2007.

[96] S.W. van den Braak and G.A.W. Vreeswijk. AVER: argument visualization for evidential reasoning. In T.M. van Engers, editor, *Proceedings of the 2006 conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference*, pages 151–156. IOS Press, Amsterdam, 2006.

[97] E.L. Rissland, K.D. Ashley, and R.P. Loui. Ai and law: a fruitful synergy. *Artificial Intelligence*, 150(1–2):1–15, November 2003.

[98] K.D. Ashley. *Modeling legal arguments: reasoning with cases and hypotheticals.* The MIT Press, 1991.

[99] V. Aleven. *Teaching case-based argumentation through a model and examples.* PhD thesis, University of Pittsburgh, 1997.

[100] H. Prakken. An exercise in formalising teleological case-based reasoning. *Artificial Intelligence and Law*, 10:113–133, 2002.

[101] T.F. Gordon. The Pleadings Game: formalizing procedural justice. In *Proceedings of the 4th international conference on Artificial Intelligence and Law*, ICAIL '93, pages 10–19. ACM, 1993.

[102] T.J.M. Bench-Capon, T. Geldard, and P.H. Leng. A method for the computational modelling of dialectical argument with dialogue games. *Artificial Intelligence and Law*, 8(2–3):233–254, September 2000.

[103] A.R. Lodder. *DiaLaw: on legal justification and dialogical models of argumentation*, volume 42 of *Law and Philosophy Library.* Kluwer Academic Publishers, 1999.

[104] D.B. Skalak and E.L. Rissland. Arguments and cases: an inevitable intertwining. *Artificial Intelligence and Law*, 1(1):3–44, March 1992.

[105] T.J.M. Bench-Capon and G. Staniford. PLAID - proactive legal assistance. In *Proceedings of the 5th international conference on Artificial Intelligence and Law*, ICAIL '95, pages 81–88, New York, NY, USA, 1995. ACM.

[106] D. Walton. *Appeal to expert opinion: arguments from authority.* Penn State University Press, University Park, PA, USA., 1997.

[107] D. Schum and P. Tillers. Marshalling evidence for adversary litigation. *Cardozo Law Review*, 13:657–704, 1991.

[108] I. Mazzotta and F. de Rosis. Artifices for persuading to improve eating habits. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 76–85, 2006.

[109] J. Aberg. Dealing with malnutrition: a meal planning system for elderly. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 1–7, 2006.

[110] R.S. Day. Comprehension of prescription drug information: overview of a research program. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 24–33, 2006.

[111] T.W. Bickmore and C.L. Sidner. Towards plan-based health behavior change counseling systems. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 14–18, 2006.

[112] P.J. Schulz and S. Rubinelli. Healthy arguments for literacy in health. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 86–95, 2006.

[113] D.L. Hunt, R.B. Haynes, S.E. Hanna, and K. Smith. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *Journal of the American Medical Association*, 280(15):1339–1346, October 1998.

[114] D.R. Sutton and J. Fox. The syntax and semantics of the PROforma guideline modeling language. *Journal of the American Medical Informatics Association*, 10(5):433–443, Sep–Oct 2003.

[115] D.W. Glasspool and J. Fox. REACT - a decision-support system for medical planning. In S. Bakken, editor, *Proceedings of the American Medical Informatics Association Symposium 2001*, page 911, 2001.

[116] J.P. Bury, C. Hurt, C. Bateman, S. Atwal, K. Riddy, J. Fox, and V. Saha. LISA: a clinical information and decision support system for collaborative care in childhood acute lymphoblastic leukaemia. In *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, page 988, 2002.

[117] D.R. Sutton, P. Taylor, and K. Earl. Evaluation of PROforma as a language for implenting medical guidelines in a practical context. *BMC Medical Informatics and Decision Making*, 6(20), 2006.

[118] R.D. Shankar, W.T. Samson, and M.A. Musen. Medical arguments in an automated health care system. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 96–104, 2006.

[119] M.K. Goldstein, B.B. Hoffman, R.W. Coleman, M.A. Musen, S.W. Tu, A. Advani, R. Shankar, and M. O'Connor. Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care. In *Proceedings of AMIA symposium*, pages 300–304, 2000.

[120] P. Tolchinsky, S. Modgil, and U. Corteś. Argument schemes and critical questions for heterogenous agents to argue over the viability of a human organ for transplantation. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 105–111, 2006.

[121] S.B. Dolins and R.E. Kero. The role of AI in building a culture of partnership between patients and providers. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 47–52, 2006.

[122] K.L. M^{c}Neil and D.S. Pimentel. Scientific discourse in three urban classrooms: the role of the teacher in engaging students in argumentation. *Science Education*, 94(5):765–793, 2010.

[123] J.F. Voss and J.A. Van Dyke. Argumentation in psychology: background comments. *Discourse Processes*, 32(2–3):89–111, 2001.

[124] M. Garcia-Mila and C. Andersen. *Argumentation in science education*, volume 35 of *Contemporary trends and issues in science education*, chapter Cognitive foundations of learning argumentation, pages 29–45. Springer, 2007.

[125] D. Cartwright and K. Atkinson. Political engagement through tools for argumentation. In P. Besnard, S. Doutre, and A. Hunter, editors, *Proceedings of the 2008 conference on Computational models of argument: Proceedings of COMMA 2007*, pages 116–127. IOS Press, 2008.

[126] M.A. Gilbert. Goals in argumentation. In D.M. Gabbay and H.J. Ohlbach, editors, *Practical Reasoning: international conference on formal and applied Practical Reasoning*. Springer Verlag, 1996.

[127] N.A. Campbell, J.B. Reece, L.A. Urry, M.L. Caine, S.A. Wasserman, P.V. Minorsky, and R.B. Jackson. *Biology*. Pearson Benjamin Cummings, 8th edition, 2008.

[128] K. Theiler. *The house mouse - atlas of embryonic development*. Springer Verlag, 1989.

[129] R. Baldock and D. Davidson. *Anatomy ontologies for bioinformatics: principles and practise*, chapter The Edinburgh Mouse Atlas, pages 249–265. Springer Verlag, 2008.

[130] M.B. Avison. *Measuring gene expression*. Taylor Francis Group, 2007.

[131] S. Venkataraman, P. Stevenson, Y. Yang, L. Richardson, N. Burton, T.P. Perry, P. Smith, R.A. Baldock, D.R. Davidson, and J.H. Christiansen. EMAGE - Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Research*, 36(1):D860–D865, 2007.

[132] C.M. Smith, J.H. Finger, T.F. Hayamizu, I.J. M^{c}Cright, J.T. Eppig, J.A. Kadin, J.E. Richardson, and M. Ringwald. The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Research*, 35(1):D618–D623, 2006.

[133] C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687–693, October 2008.

[134] E. Antezana, M. Kuiper, and V. Mironov. Biological knowledge management: the emerging role of the semantic web technologies. *Briefings in bioinformatics*, 10(4):392–407, May 2009.

[135] K. M^{c}Leod and A. Burger. Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In N. Guimaraes and P. Isaís, editors, *Proceedings of IADIS International Conference Applied Computing*, pages 489–492, Salamanca, Spain, Februrary 2007. IADIS Press.

[136] P. Ziegler and K.R. Dittrich. Data integration — problems, approaches, and perspectives. In J. Krogstie, A.L. Opdahl, and S. Brinkkemper, editors, *Conceptual modelling in information systems engineering*, pages 39–58. Springer Verlag, Berlin, 2007.

[137] V.S. Subrahmanian and L. Amgoud. A general framework for reasoning about inconsistency. In *Proceedings of the 20th international joint conference on artifical intelligence*, IJCAI'07, pages 559–604. Morgan Kaufmann Publishers, 2007.

[138] M.V. Martinez, F. Parisi, A. Pugliese, G.I. Simari, and V.S. Subrahmanian. Inconsistency management policies. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 11th international conference, KR, 2008*, pages 367–377, 2008.

[139] M.V. Martinez and A. Hunter. Incorporating classical logic argumentation into policy-based inconsistency management in relational databases. In *The uses of computational argumentation*, AAAI Fall Symposium. Technical Report FS-09-06, pages 52–57. AAAI Press, 2009.

[140] I. Ben-Gai. *Encyclopedia of statistics in quality and reliability*, chapter Bayesian networks. John Wiley & Sons, 2008.

[141] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyse expression data. *Journal of Computational Biology*, 7(3–4):601–620, 2000.

[142] H. Irandoust. Attitudes for achieving user acceptance: explaining, arguing, critiquing. In *In proceedings of the 7th international command and control research and technology symposium*, Quebec City, Canada, 2002.

[143] D. Walton. Can argumentation help AI to understand explanation. *Kunstliche Intelligenz*, 22(2):8–12, 2008.

[144] M.R. Wick. Expert system explanation in retrospect: a case study in the evolution of expert system explanation. *Journal of Systems and Software*, 19(2):159–169, October 1992.

[145] A. Stranieri and J. Zeleznikow. A survey of argumentation structures for intelligent decision support. In *Proceedings of the 5th international conference of the society for decision support systems (ISDSS'99)*, pages 18–21. Monash University Press, 1999.

[146] S.J. Alvarado. *Understanding editorial text: a computer model of argument comprehension.* Kluwer Academic Publishers, 1990.

[147] T.J.M. Bench-Capon, D. Lowes, and A.M. M^{c}Enery. Argument-based explanation of logic programs. *Knowledge-Based Systems*, 4(3):177–183, September 1991.

[148] L.R. Ye. The value of explanation in expert systems for auditing: an experimental investigation. *Expert Systems with Applications*, 9(4):543–556, 1995.

[149] L.R. Ye and P.E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Q*, 19(2):157–172, June 1995.

[150] The ASPIC Consortium. Deliverable d2.1 - theoretical framework for argumentation. www.argumentation.org, July 2004.

[151] I. Bratko. *PROLOG Programming for Artificial Intelligence.* Addison Wesley, 2000.

[152] D.A. Grant and E.A. Berg. A behavioural analysis of degree or reinforcement and ease of shifting to new responses in a weigi-type card sorting problem. *Journal of Experimental Psychology*, 38:404–411, 1948.

[153] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[154] The ASPIC Consortium. Deliverable d1.4 - final consensus framework of argumentation concepts. www.argumentation.org, November 2005.

[155] J. Fox, E. Black, D.W. Glasspool, S. Modgil, A. Oettinger, V. Patkar, and M. Williams. Towards a general model for argument services. In T.W. Bickmore and N. Green, editors, *AAAI Spring Symposium: Argumentation for consumers of healthcare*, pages 52–57, 2006.

[156] M. Caminada and L. Amgoud. An axiomatic account of formal argumentation. In *Proceedings of the 20th national conference on artificial intelligence*, volume 2 of *AAAI'05*, pages 608–613. AAAI Press, 2005.

[157] The ASPIC Consortium. Deliverable d2.6 - final review and report on formal argumentation system. www.argumentation.org, January 2006.

[158] K. Sutherland, K. M^{c}Leod, and A. Burger. Semantically linking web pages to web services in bioinformatics. In *3rd International AST Workshop*, 2008.

[159] G. Ferguson, K. M^{c}Leod, K. Sutherland, and A. Burger. Sealife evaluation. Technical Report 0063, Dept of Computer Science, Heriot-Watt University, 2009.

[160] B. Shneiderman. *Designing the user interface: strategies for effective human-computer interaction.* Addison-Wesly, 2nd edition, 1992.

[161] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

[162] B.M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5–6):527–539, November/December 1987.

[163] H. Atoyan, J. Duquet, and J. Robert. Trust in new decision aid systems. In *Proceedings of the 18th International Conference of the Association Fracophone d'Interaction Homme-Machine*, IHM'06, pages 115–122. ACM, 2006.

[164] D.D. Suthers and C.D. Hundhausen. An experimental study of the effects of representational guidance on collaborative learning processes. *Journal of the Learning Sciences*, 12(2):183–218, April 2003.

[165] D.D. Suthers, R. Vatrapu, R. Medina, S. Josepth, and N. Dwyer. Beyond threaded discussion: representational guidance in asynchronous collaborative learning environments. *Computers and Education*, 50(4):1103–1127, May 2008.

[166] E.M. Nussbaum, D.L. Winsor, Y.M. Aqui, and A.M. Poliquin. Putting the pieces together: online argumentation vee diagrams enhance thinking during discussions. *International Journal of Computer-Supported Collaborative Learning*, 2(4):479–500, 2007.

[167] S. M^{c}Alister, A. Ravenscroft, and E. Scanion. Combining interaction and context design to support collaborative argumentation using a tool for synchronous CMC. *Journal of Computer Assisted Learning*, 20(3):194–204, June 2004.

[168] B.B. Schwarz and A. Glassner. *Arguing to learn: confronting cognitions in computer-supported collaborative learning evironments*, chapter The blind and the paralytic: fostering argumentation in social and scientific domains, pages 227–260. Kluwer Academic Publishers, 2003.

[169] K. Stegmann, A. Weinberger, and F. Fischer. Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, 2(4):421–447, 2007.

[170] O. Scheuer, F. Loll, N. Pinkwart, and B.M. McLaren. Computer-supported argumentation: a review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102, March 2010.

[171] B. Verheij. Artificial argument assistants for defeasible argumentation. *Artificial Intelligence*, 150(1–2):291–324, November 2003.

[172] D. Gaertner and F. Toni. Computing arguments and attacks in assumption-based argumentation. *IEEE Intelligent Systems*, 22(6):24–33, Nov.-Dec. 2007.

[173] M. Morge and P. Mancarella. The hedgehog and the fox: an argumentation-based decision support system. In *Proceedings of the 4th international conference on argumentation in multi-agent systems*, ArgMAS'07, pages 114–131. Springer-Verlag, May 2008.

[174] G.A.W. Vreeswijk. Argumentation in bayesian belief networks. In *Proceedings of the First international conference on Argumentation in Multi-Agent Systems*, ArgMAS'04, pages 111–129. Springer-Verlag, 2005.

[175] I. Mazzotta, F. de Rosis, and V. Carofiglio. PORTIA: a user-adapted persuasion system in the healthy eating domain. *IEEE Intelligent Systems*, 22(6):42–51, November 2007.

[176] G.A. Johnson, A. Badea, J. Brandenburg, G. Cofer, B. Fubara, S. Liu, and J. Nissanov. Waxholm Space: an image-based reference for coordinating mouse brain research. *NeuroImage*, 53(2):365–372, November 2010.

[177] L.A.L. Silva, B.F. Buxton, and J.A. Campbell. Enhanced case-based reasoning through use of argumentation and numerical taxonomy. In D. Wilson and G. Sutcliffe, editors, *Proceedings of the 20th international Florida artificial intelligence research society conference*, pages 423–428. AAAI Press, May 2007.

[178] F.M. Shipman III and C.C. Marshall. Formality considered harmful: experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer-Supported Cooperative Work*, 8(4):332–352, 1999.

[179] M. Polanyi. *The tacit dimension.* DoubleDay & Company, 1966.

[180] J. Bliss. *Learning with artificial worlds: computer based modelling in the curriculum*, chapter From mental models to modelling. The Falmer Press, 1994.

[181] R.M. Schumacher and M.P. Czerwinski. *The psychology of expertise: cognitive research and empirical AI*, chapter Mental models and the aquisition of expert knowledge, pages 61–79. Lawrence Erlbaum Associates, 1992.

[182] I.M. Greca and M.A. Moreira. Mental models, conceptual models and modeling. *International Journal of Science Education*, 22(1):1–11(11), January 2000.

[183] S. Mittal and C.L. Dym. Knowledge acquisition from multiple experts. *AI Magazine*, 6(2):32–36, 1985.

[184] H. Lindgren. Towards using argumentation schemes and critical questions for supportng diagnostic reasoning in the dementia domain. In *Proceedings of Computational Models of Natural Aguments*, CMNA'09, pages 10–14, 2009.

[185] H. Lindgren and P. Winnberg. Evaluation of a semantic web application for collaborative knowledge building in the dementia domain. In M. Szomszor and P. Kostkova, editors, *Electronic Healthcare - 3rd international conference, eHealth 2010, revised selected papers*, volume 69 of *Lecture Notes of the Institute for Computer Science, Social Informatics and Telecommunications Engineering*, pages 62–69. Springer, 2010.

[186] J. Fox and S. Parsons. Arguing about beliefs and actions. In *Applications of Uncertainty Formalisms*, pages 266–302. Springer Verlag, 1998.

[187] A.S. Coulson, D.W. Glasspool, J. Fox, and J. Emery. RAGs: a novel approach to computerized genetic risk assessment and decision support from pedigrees. *Methods of Information in Medicine*, 40(4):315–322, 2001.

[188] A.H. Murphy and R.L. Winkler. Probability forecasting in meterology. *Journal of the American Statistical Association*, 79(387):489–500, September 1984.

[189] A.E. Hoerl and H.K. Fallin. Reliability of subjective evaluations in high incentive situation. *Journal of the Royal Statistical Society. Series A (General)*, 137(2):227–230, 1974.

[190] J.J. Christensen-Szalanski and J.B. Bushyhead. Physician's use of probabilistic information in a real clinic setting. *Journal of Experimental Psychology, Human Perception and Performance*, 7(4):928–935, August 1981.

[191] M. Hynes and E. Vanmarcke. Reliability of embankment performance predictions. In *Proceedings of the ASCE Engineering Mechanism Division, Speciality Conference*. University of Waterloo Press, 1976.

[192] S.O. Hansson. Decision theory a brief introduction. http://home.abe.kth.se/ soh/decisiontheory.pdf, 2005.

[193] H. Lindgren. Conceptual model of activity as tool for developing a dementia care support system. In M. Ackerman, R. Dieng-Kuntz, C. Simone, and V. Wulf, editors, *Knowledge management in action*, volume 270, pages 97–109. Springer, 2008.

[194] K. Wolstencroft, A. Brass, I. Horrocks, P. Lord, U. Sattler, D. Turi, and R. Stevens. A little semantic web goes a long way in biology. In Yolanda Gil, Enrico Motta, V. Benjamins, and Mark Musen, editors, *The Semantic Web – ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 786–800. Springer Berlin / Heidelberg, 2005.

[195] N. Jiménez-Lozano, J. Segura, J. Macías, J. Vega, and J. M. Carazo. aGEM: an integrative system for analyzing spatial-temporal gene-expression information. *Bioinformatics*, 25(19):2566–2572, October 2009.