Using Argumentation to resolve conflict in Biological Databases

Kenneth M^{c}Leod
kcm1@hw.ac.uk
Dept. of Computer Science
Heriot-Watt University

Gus Ferguson
gus.ferguson@gmail.com
Dept. of Computer Science
Heriot-Watt University

Albert Burger
ab@macs.hw.ac.uk
Heriot-Watt University, and
MRC Human Genetics Unit

Abstract

Biological experiments are published in a series of online databases. Due to the inherent complexity of these experiments, their conclusions can be contradictory. This causes the databases to be inconsistent, which creates problems for human users. Previous work looked at creating a system based on Argumentation to resolve this issue. This work presents an evaluation of that system, focusing on the effectiveness of its presentation of arguments and the result of the argumentation process. The work shows that the system can be understood by users, even though their presentation preferences vary substantially.

1 Introduction

Biologists have access to an ever increasing number and range of online data resources [Galperin and Cochrane, 2009]. Many of these resources contain inconsistent data. This is not surprising as Biology is a complex science in which countless parameters affect the outcome of every experiment. Added to this is the human element that causes two identical results to be evaluated differently by different people. The consequence is that two seemingly identical experiments can produce contradictory outcomes. These experiments may be stored in one or more of the online resources that service a particular field.

If both of these experiments are published by the same resource, it becomes inconsistent. However, if each experiment is published by a different resource, then the inconsistency is between resources and becomes harder to detect. Regardless of where it occurs, inconsistency confuses users, forcing them to research further in order to answer their query.

In [M^{c}Leod and Burger, 2007] it was suggested that argumentation could be one solution to this problem. By using all the resources in a field, arguments can be created for and against potential answers to a query. These arguments can be presented to users, providing them with a powerful set of knowledge that could be used to identify the most likely solution to the query. [M^{c}Leod and Burger, 2008] described the implementation of this idea.

This paper describes the presentation of such arguments for biological users, and the evaluation of this presentation. It starts with a description of the biological domain the work is set in (Section 2), proceeds in Section 3 with a summary of the previous work, before Section 4 describes the evaluation undertaken and its results. These results point to the need for an evaluation of the best way to visually present arguments, and this is discussed in Section 5 before conclusions are provided in Section 6.

2 Biology

Commonly in Biology, once an experiment is published in a suitable peer reviewed journal, its result is published in an online database. In 2008 the domain of Molecular Biology created ninety-five new databases and revised eighty-six existing ones [Galperin and Cochrane, 2009] - this gives an indication of the importance and prevalence of these resources. Normally they will store some basic provenance (e.g. the names of the researchers, and a link to the published paper) along with basic details of the experiment and its conclusion. The exact information stored varies according to the database. Many of these databases are edited by humans, so-called *manual curation*. If manually curated the database may also have some extra human generated annotation.

2.1 Gene Expression Information

This work focuses on one specific sub-domain of Molecular Biology - gene expression for the developmental mouse. *Genes* are effectively the instructions that tell the body what to build (e.g. one set of genes results in a nose, and a second set a thumb, et-cetera) and how to function (e.g. detect scents). There are various levels of gene expression ranging from *strong* through to *not expressed*. Genes that are not expressed play no role in the creation or function of the tissue. Other levels of expression imply the genes have some role, these genes are said to be *expressed* in the tissue.

In addition to playing a key role in the development of regular features, genes are active in the creation of some abnormal features, such as cleft lips, and diseases such as cancer. Consequently, biologists study the level of gene expression in regular tissues and compare the results to those from abnormal tissues. Experimentation on human subjects is regarded as unacceptable, so the scientists use a range of model organisms, such as the mouse and zebrafish. Each model organism has its own research community, and each community has

access to at least one online database that publishes their experimental results.

Broadly speaking, there are two forms of gene expression experiment: in-situ and microarray based. The latter places the emphasis on gathering the quantity of each gene expressed, at the expense of some accuracy in terms of where it is expressed. In-situ does the opposite. It produces an image of the subject, with the areas in which a gene is expressed vividly coloured, thus allowing spatial processing to identify exactly where the gene is expressed.

The *where* refers to the actual tissue. If dealing with an adult subject, this is a 3-D location. However, developmental subjects are also used - the term *developmental* being used to describe a subject that is somewhere between conception and birth. This extra element of time results in a 4-D mapping.

The subject of this work, is the developmental mouse. Its development, from conception to birth, is split into twenty-six stages, called Theiler Stages. For each stage there is an anatomical ontology. A collection of 3D models (one for each stage), the anatomy ontologies, and the links between them make up the Edinburgh Mouse Atlas Project (EMAP) [Baldock and Davidson, 2008]. Gene expression results are mapped to one or more of the tissues in these 26 ontologies as well as the 3D models.

In addition, resources provide some indication of the quantity of the gene expressed in the tissue. For microarray methods this is a number; however, the in-situ experiments are less precise, often only providing a natural language description e.g. strong or weak.

2.2 Inconsistency and Contradiction

Many of the online databases mentioned previously, feature inconsistent and incomplete information. For example, in the field of gene expression, few databases hold information on every gene in every structure of an organism. In order to reduce the gaps in knowledge, scientists continually research new methods and technologies to help them work faster and more accurately. This may create a need to repeat some experiments with the new method/technology, temporarily increasing the volume of information needed to create a complete model of the domain.

In addition to incomplete information the researchers must also consider conflicting information. Two research groups may conduct seemingly similar experiments, but obtain different results and conclusions. This may be due to experimental error or a slight variation in experimental conditions. Despite the variations, both results will be published by journals and subsequently entered into (possibly the same) online databases. Therefore the distributed online databases contradict each other, and themselves.

Consider a gene expression database for the developmental mouse called GXD [Smith *et al.*, 2007]. A query that asks which genes are not expressed in the Brain in Theiler Stage 24 (TS24) will include in the answer the gene *Tnc*. This is because one experiment provided this result. However a more careful examination will reveal that there are fourteen experiments with the opposite result. Clearly any work based solely on the first query is likely to be suspect.

Often the correct conclusion is not so obvious. A second resource publishing gene expression information for the developmental mouse, EMAGE [Venkataraman *et al.*, 2008], has one experiment suggesting that the gene *Hoxb1* is expressed in the Neural Ectoderm TS11, and one experiment suggesting it is not. Biologists generally regard the two levels of gene expression as mutually exclusive, so a resolution is required. Making a decision requires the time to read the underlying research papers and the expertise to understand them.

These issues are not unique to EMAGE and GXD - they are simply used to illustrate that these issues need to be considered when using online resources publishing biological information.

2.3 Use Case Databases

The work described in [M^{c}Leod and Burger, 2008] focuses on EMAGE and GXD; however, it suggested the need for the inclusion of further databases such as CGAP [Strausberg, 1999].

EMAGE and GXD are considered as complementary. These resources publish the same type of information, and do so using the same anatomy ontology (EMAP). Despite the similarities, differences exist. Firstly, although the data overlaps the resources are not identical - some experiments published in EMAGE are not available in GXD and vice versa. Furthermore, only EMAGE creates a distinction between *Textual Annotations* and *Spatial Annotations*. The results of in-situ gene expression experiments (2D section images - Figure 1 contains an example at the right of the image) can be described with respect to the EMAP anatomy ontology or spatially mapped into EMAGE's 3D embryo models (one per Theiler Stage) of EMAP. These are referred to as Textual Annotation and Spatial Annotation, respectively. GXD features only results mapped to the EMAP anatomy ontology.

CGAP is a database associated with cancer genes in the mouse and human. A subset of its data relates to gene expression for the developmental mouse, in particular one class of microarray technique called SAGE. Unlike the previous resources, this data is not explicitly tied to the EMAP anatomy ontology. Instead the researchers have chosen to use their own anatomy ontology. Unfortunately, there is no direct one-to-one mapping between the EMAP and CGAP anatomy ontologies, so only subsets of the data can be used. These subsets correspond to individual expert created mappings between the ontologies. A further difference between this database and the other two, is that it provides a count of the number of genes expressed in a tissue as opposed to the natural language value given by EMAGE and GXD.

3 Using Argumentation with EMAGE and GXD

Biologists, in general, do not have a background in formal logic. However, they do have a desire to understand the processes with which they are being given new information. Argumentation, with its clear mapping to real world communication, seemed a good choice for reasoning over the data in

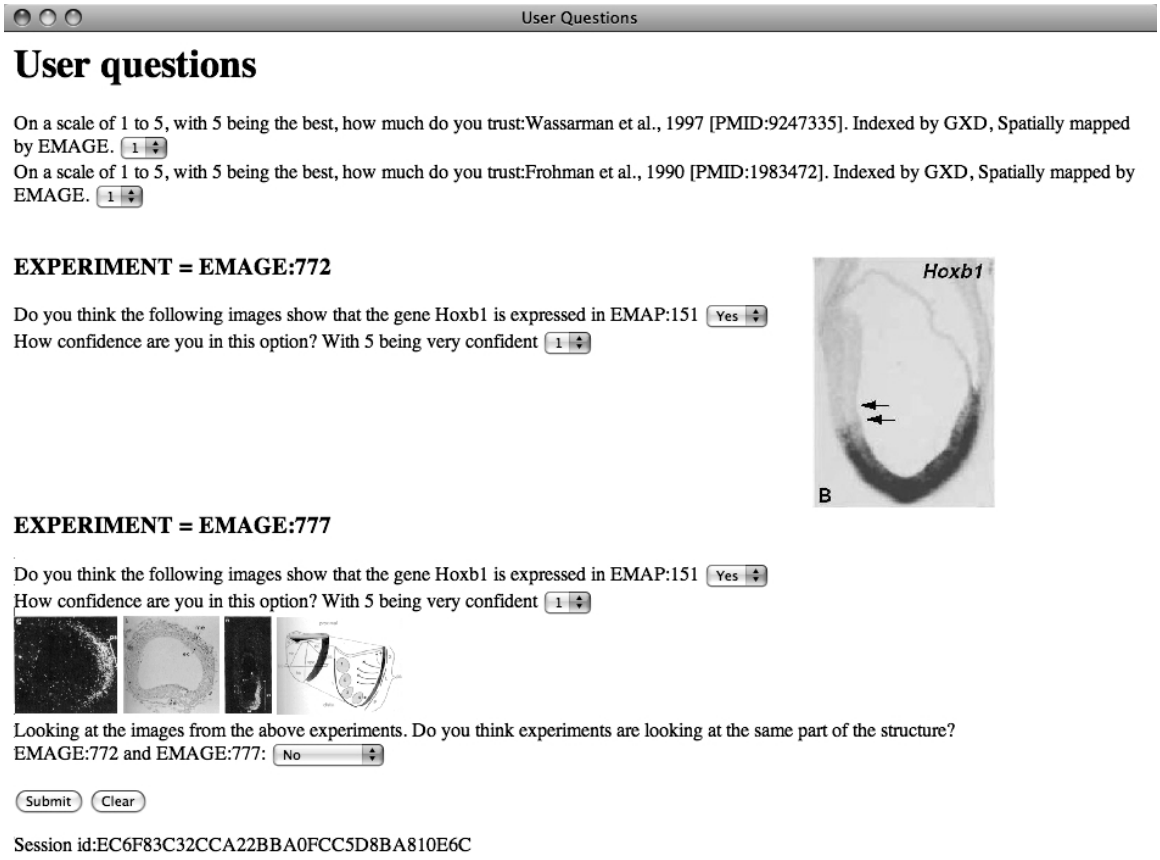

Figure 1: Screenshot from implemented system - asks the user for their opinion of the researchers and experimental images

EMAGE and GXD. An opinion strengthened by its deployment in the related world of medicine [Bickmore and Green, 2006]. Furthermore, this is not the first time argumentation has been employed in Biology, with [Jefferys *et al.*, 2006] using it to evaluate the output of a predictive tool. Although the purpose of that work is different from ours, it suggests the underlying suitability of argumentation for professionals working in the life sciences.

Argumentation has been used within this work to resolve inconsistencies across biological data resources. A variety of other mechanisms to integrate data and resolve inconsistency exist. For example, data reconciliation (a.k.a. data fusion) uses a function to turn multiple possible values into a single value, e.g. computing the average of four numbers (e.g. [Motro and Rakov, 1998]). A second possible mechanism would create multiple query plans for the resources, then select the best according to information quality criteria (e.g. [Naumann, 1996]). This work is not an attempt to replace these mechanisms. It is not concerned with automatically resolving conflict, but instead wishes to determine whether or not argumentation can enable biologists to resolve the differences themselves.

The argumentation engine used in this project was created as part of the ASPIC project [Fox *et al.*, 2007] . This engine takes domain information, and knowledge of how to interpret that information (as a series of inference rules) and uses them to create arguments by backward chaining through the rules in response to a query from the user.

In order to provide the expert knowledge vital for this work, a curator of the EMAGE database was recruited. Although the argumentation engine's inference rules are written in first order predicate logic, communication with the expert is restricted to natural language. There was a need to provide some degree of formality in order to correctly structure the rules and accurately record the exceptions and other forms of attack. Thus the basic notion of *Argument Schemes* [Walton *et al.*, 2008] was employed as a bridge between the expert and argumentation engine, with the transformation to logical inference rules based on the work of Verheij [Verheij, 2003].

3.1 Schemes

It should be noted, that the schemes used in this work are a subset of those that would be used in a fully comprehensive system. To produce such a system for developmental mouse gene expression, one would have to include all resources providing mouse gene expression information, particularly those that provide different types of microarray data (CGAP provides only one of many). Furthermore, other ways of obtaining gene expression results would need to be considered. For example one might consider which genes are expressed in the

equivalent tissue in the zebrafish, or what information could be obtained from Biological Pathways (a network of chemical reactions in a particular system or process, e.g. metabolism). The resulting explosion of resources and schemes is far beyond the scope of any single project.

The schemes suggested by the expert for this work are of two types. The first relating to the user's confidence in resources (e.g. EMAGE), journals, individual researchers, and techniques (e.g. Spatial Annotations). The second category is for broadly accepted inferences, e.g. Textual Annotations are generally more reliable than Spatial Annotations.

User confidence in the technique of Spatial Annotations

The schemes for confidence were often not associated with critical questions, because the expert believed that there was no need for them. If someone did not have trust in something, this was a perfectly valid viewpoint that need not be explored further. An example is provided below.

Result R is based on a Spatial Annotation.
The user has low confidence in Spatial Annotations.
If the user has C confidence in Spatial Annotations, then they have C confidence in R.
Therefore, the user has low confidence in R.

Textual Annotations and Spatial Annotations

The intuition behind the following scheme is that generally Textual Annotations are more reliable than the Spatial Annotations. However, if the researcher created or approved the Spatial Annotation this may not hold. This is also true if the Textual Annotation seems unreliable, or other experiments/resources agree with the Spatial Annotation.

Result $R1$ came from a Textual Annotation.
Result $R2$ came from a Spatial Annotation.
Textual Annotations may be more accurate than Spatial Annotations.
Therefore, $R1$ may be more accurate than $R2$.

Questions:

1. Who created the Spatial Annotation?
2. Who approved the Spatial Annotation?
3. Is there genuine conflict between the Spatial and Textual Annotations?
4. Is the Textual Annotation trustworthy?
5. Which annotations can be obtained from other experiments?
6. What conclusions can be drawn from other resources?

3.2 Implementation

In addition to providing the schemes, the expert was asked to arrange them in order of importance, the most important being the ones in which he had the most confidence. Following this, scores were assigned to the schemes. These scores were associated with the rules and used to determine the strength of arguments and thus resolve conflict between them.

With the schemes complete, clients were programmed to take advantage of the programmatic interfaces provided by EMAGE and GXD. When a user specifies a gene and structure in which (s)he is interested, the clients pull the relevant data and convert it for use within the argumentation engine. The system's user interface asks the user their confidence levels (e.g. in researchers who conducted relevant experiments - a screenshot of this can be seen in Figure 1) and presents this information to the ASPIC argumentation engine.

When the domain data and expert knowledge is loaded into the argumentation engine's knowledge base, a query (*Is the gene expressed in the structure?*) can be sent to the system and the resulting arguments displayed to the user. A simplified architecture of the system can be seen in Figure 2.

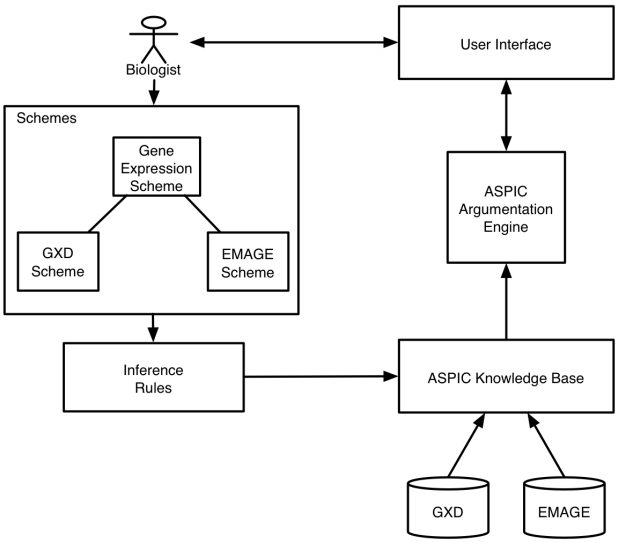

Figure 2: Simplified architecture of system

With a simple text based user interface, run as a series of JAVA Servlets, an initial expert evaluation of the system was conducted with the development team, including the biological expert. This ensured that the system generated the kind of arguments the expert wished to see, and showed that the arguments contained the information necessary for him to make a decision. During this evaluation two issues were identified: i. the need for a proper graphical user interface (GUI), and ii. the importance of including other databases in addition to EMAGE and GXD.

More details on the above can be found in [M^{c}Leod and Burger, 2008].

Subsequently, a third source, CGAP was included in the system, and a prototype GUI was designed and implemented. Expert users indicated they would prefer a textual representation of the actual arguments. In addition a visual summary of the arguments' relationships to one another could also prove useful as the system creates a number of arguments, both for and against the gene being expressed.

The example in Figure 3 presents the arguments for the query: *Is the gene Bmp4 expressed in the Telencephalon in TS15?* Here arguments are represented by circles and identi-

fied by letters (these letters are connected to textual representations of the arguments, e.g. Figure 4). The argument's claim is distinguished by an arrow leaving the circle and pointing at a box - one representing the gene being *expressed* and the other *not expressed*. The strength of the argument is captured by the line - the strongest argument (as identified by the argumentation engine from the scores assigned to the schemes) has a solid line, the weaker arguments a dashed line and the weakest a dotted line. The conclusion, as determined by the strongest argument, is given by an arrow leaving a box (expressed or not expressed) and heading towards a box containing the word *succeeds* (in the actual system the lines for the arguments were coloured blue in order to distinguish them). Consequently, Figure 3 suggests that *Bmp4* is expressed in the Telencephalon in TS15 (which has EMAP ontology ID *EMAP:1212*).

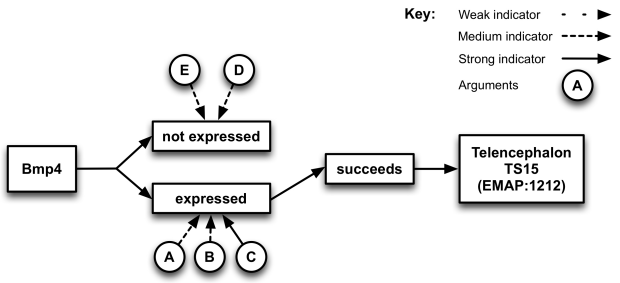

Figure 3: Visual summary of arguments produced by system - arguments are linked to textual representations, e.g. Figure 4

Bmp4 is not expressed in TELENCEPHALON TS15. 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=79%

Figure 4: Textual representation of an argument

Prior to full implementation and linking of the GUI to the rest of the system, an evaluation was undertaken, focusing on the use of visualisation and natural language representations to communicate information.

4 Evaluation

Standard evaluation methods and protocols were used alongside bespoke protocols to evaluate and obtain feedback from users of the prototype GUI. The evaluation aimed to obtain both summative and formative evaluation of the system's functionality and evaluate specific aspects of the conceptual design of data presentation and visualisation.

A group of eighteen users were recruited for the evaluation. Ten perform various roles, from biological database curators to system developers, on the Edinburgh Mouse Atlas Project at the Human Genetics Unit (HGU) of the Medical Research Council in Edinburgh (http://genex.hgu.mrc.ac.uk). This group included expert users who had previously provided opinion on aspects of the system during the development. The other eight users are Computer and Life Sciences students at Heriot-Watt University. The evaluation was conducted using an Apple Macintosh computer in a dedicated room in two sites at Heriot-Watt and the HGU.

Test scenarios were designed based on typical tasks the system is intended to perform with the data for these tasks generated using the argumentation system. This data was hard-coded into the GUI to forestall issues of availability of online resources, and network performance variability between the two locations.

The first argumentation scenario consisted of a walkthrough of the system using an example relating to expression of a gene in the developmental mouse brain, with the default settings. The user was then presented with a graphical and textual representation of the outcome. Users were asked specific questions regarding the process and presentation of the results.

The second scenario involved the user walking through the same process, but altering selections for the databases used and the level of trust they placed in some researchers and journals (a screenshot of the GUI used to obtain this information is given in Figure 5). The users were then presented with results modified appropriately to the altered parameters, and were again asked specific questions regarding the results and their presentation. Finally, users were asked specific questions regarding their understanding of and views on, argumentation, the process of the argumentation system, and the presentation of the results.

Two evaluators were used, one interacting 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 procedure and users were prompted to comment, ask questions or ask for help freely at any stage during the evaluation.

A standard consent form including a brief explanation of the system and its purposes was used. Users then filled in a background questionnaire on education/training, work experience and familiarity with appropriate bioinformatics online resources and journals. They then completed the scenarios and specific questionnaires before finishing with a general system usability questionnaire, modified from Shneiderman's QUIS questionnaire [Shneiderman, 1992].

Further details of the evaluation, and the protocols used during it, can be seen in [Ferguson *et al.*, 2009].

4.1 Results and Discussion

When analysing the results, the eighteen users were split into two groups of nine, a Biologist group comprising users with biological training and experience, and a non-biologist group for users with little or no biological training or experience. The first group was split further into four users considered to be experts in gene expression and the resources used in this study (EMAGE, GXD, and CGAP), and the remaining five users who have no specialist knowledge of gene expression. The evaluation focused on user responses in three distinct areas: i. the GUI and the system, ii. the presentation of the arguments, and iii. the display of results.

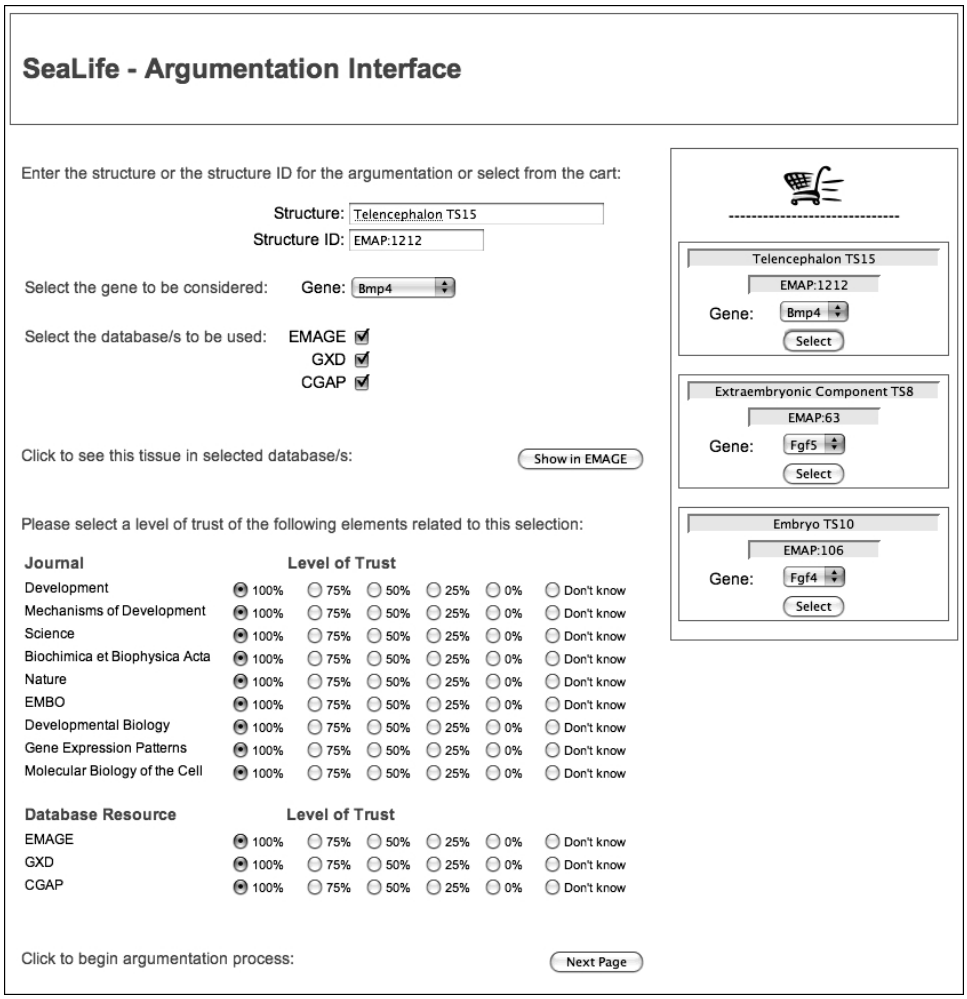

Figure 5: Screenshot of the prototype GUI - start page asking the user which gene-tissue pair they are interested in, and then for their confidence in a variety of journals and online resources

Evaluation of the system

Users performed both argumentation scenarios, the first walkthrough with default settings and the second with the user lowering confidence levels in a particular journal and a particular researcher. They were then asked questions aimed at determining their understanding of the argumentation process and results.

In both scenarios the majority of the responses of the non-biologist group showed correct understanding of the process and results. However, the responses of more than half (56%) of the biologist group indicated that they had used their own knowledge of the data to over-ride the argumentation system results, particularly in the second scenario. Here comments show that the expert biologists did not accept the changes in trust status for the journal and the author arbitrarily imposed by the evaluation scenario. On these grounds the expert biologists either did not answer the question or expressed their disagreement with the system.

Following the walkthroughs, users were asked a number of questions about their experience recording their responses from 1 to 9 on a Likert scale [Likert, 1932]. Regarding the amount of information presented by the argumentation system (1 - Too little to 9 - Too much), fifteen users (83%) rated it in the 4-6 range, with twelve (67%) scoring it at 5 (just right).

On how well they understood the arguments (1 - Not at all to 9 - Completely), the ratings by the non-biologists (median = 3) were significantly lower than those of the biologists (median = 7) ($p = 0.0121$). Comments and observations indicate that most of the problems the biologist group encountered were related to the confidence ratings for the arguments. The low levels of understanding among the non-biologists appeared to be largely due to lack of background biological knowledge.

A reasonable conclusion to draw from this is that the presentation of the results from the argumentation system is clear and allows users, who are merely using the system as an expert system, to reach correct solutions. However, experts and users experienced in the field have issues with trusting the argumentation system and the results it produces.

The issues related to user trust in decision aid systems have been widely discussed since Muir's work in the area in the late 1980s [Muir, 1987]. The spreading of distrust to the rest of the system as described by Muir would not have occurred in this instance, as the participants in the evaluation were fully aware of the contrived nature of the parameter changes in the scenario. Work on user trust in systems has revolved largely around modelling and measuring trust [Atoyan *et al.*, 2006] and there has been no work that establishes typical levels of trust of experts in decision aid systems in general. Further work would be needed to reliably establish levels of trust for the argumentation system.

Evaluation of arguments

Early in the development process, the expert user had suggested that he would prefer arguments to be presented as a natural language paragraph. He explicitly ruled out the use of bullet points, and the "show/hide" of detailed information, preferring all information to be immediately available. With a view to the non-expert users, it was decided that a graphical form of representation should be compared with the textual form during the evaluation.

Many systems, such as Araucaria [Reed and Rowe, 2001] and Carneades [Gordon, 2007], visualise arguments in a bottom-up graph with the conclusion at the top. This was chosen as the form for the graphical representation for the evaluation. The same argument was presented in a textual form (Figure 4), and then a graphical form (Figure 6). The text used was the same in both representations; however, the graphical form made explicit the structure of the argument and the links between the different components. The argument is based on the intuition that if a gene is not found in a tissue, it cannot be found in any subpart of that tissue (propagation). Consequently, the argument concludes that *Bmp4* is not expressed in the Telecenphalon in Theiler Stage 15.

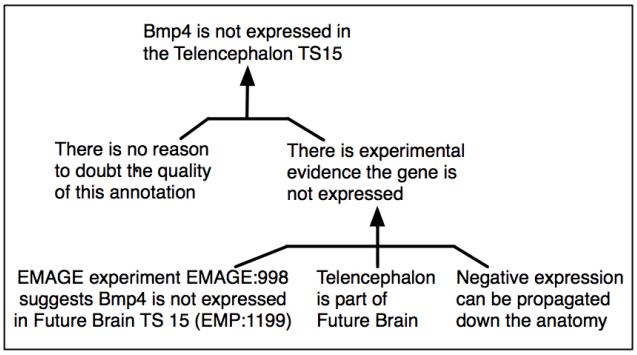

Figure 6: Graphical representation of argument in Figure 4

The results from the evaluation show that the expert biologists were evenly split with two preferring the textual and two the graphical representations of the arguments. Four of the other biologists preferred the graphical representation while one was undecided. Of the non-biologists, four preferred the textual representation, three the graphical and one was undecided. (Totals: Graphical - nine; Textual - six; and, Undecided - two). This indicates the need for both representations to be presented by the system.

Asked how the representations could be improved, a number of users commented that they felt the visual representation was upside down, and the premises should be at the top with the conclusion at the bottom. This fed into a study of the graphical presentation discussed in Section 5. Further comments included the need for an increase in the amount of explanation given in both representations.

Evaluation of the presentation of the results

In the final body of the evaluation, the visual technique used to summarise the arguments (see Figure 3) was tested to ensure it was understandable. Users were asked to identify the strength (strong, moderate, or weak) of three arguments contained in the diagram. In total fifty-four answers were given (three arguments multiplied by eighteen users). Forty-seven of these were correct (87%). Apart from two outliers, most of the users found the diagram very easy to understand.

In the evaluation scenarios, the system's result page was headed by a single line summary representing the conclusion the system had drawn, followed by the summary diagram, and then the textual representation of all the arguments. Users were asked which sections they used to reach their decision on whether or not the gene was expressed. The results are summarized as follows:

- six users used all three elements;
- three used the summary and the diagram;
- three used the diagram only;
- three used the diagram and the detailed descriptions of the arguments;
- two used the detailed descriptions of the arguments only;
- three used the summary and the detailed descriptions of the arguments;
- and, no users relied only on the summary.

Focusing exclusively on the biological group, only five out of nine used the summary, whereas seven out of nine used the diagram, and a further seven out of nine the textual arguments.

Overall, this suggests that all three elements are employed by the users in reaching their decision on the results of the argumentation.

5 Evaluation of argument graphs

User feedback identified a need to determine if the tree-based argument graph should be displayed bottom-up (i.e. with the conclusion at the top as in Figure 7) or top-down (i.e. with the conclusion at the bottom as in Figure 8). Another form of argument presentation, reading from left-to-right, was proposed by Toulmin [Toulmin, 2003]. It was decided to include a simplified form of this presentation in the evaluation (e.g. Figure 9). A similar argument to that used in the previous evaluation appears in each graph with some amendments to clarify the explanation.

The evaluation of the representations was undertaken as an online survey, with a user group drawn from biologists and bioinformaticians, recruited by email invitation through

the Scottish Bioinformatics Forum (www.sbforum.org) mailing list, and staff from the HGU. The survey used the three graphs shown in Figures 7 to 9. The first stage required participants to select their preference between the bottom-up and the top-down versions of the tree-based argument graph. The second stage 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.

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, while only one (14%) of the seven users who initially preferred the bottom-up version choose the Toulmin-like graph as their overall preferred representation.

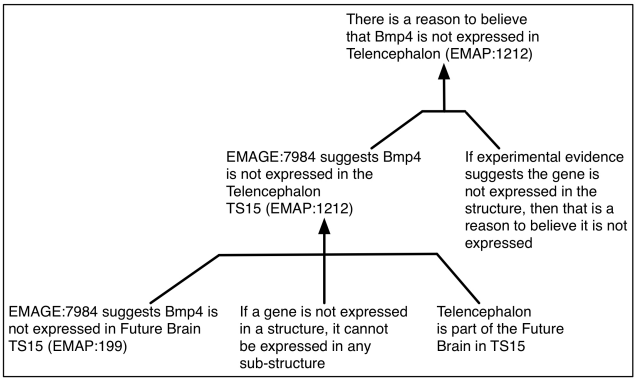

Figure 7: Bottom-up version of argument

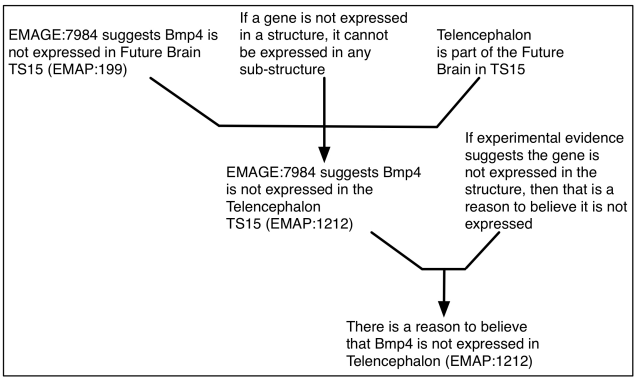

Figure 8: Top-down version of argument

6 Conclusion

Biology is fundamentally complex, which can cause seemingly identical experiments to be very slightly different.

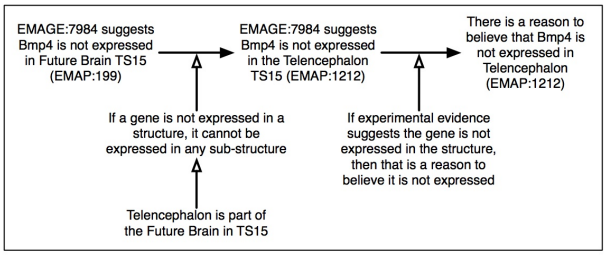

Figure 9: Toulmin-like version of argument

Therefore two experiments can produce two different results when it seems they should be identical. This inconsistency is propagated through the peer-reviewed journals into online databases that publish the results. In addition to these databases being inconsistent, there is the problem of inconsistency across databases, as many fields are served by more than one.

The system discussed in this paper attempts to address this issue using argumentation, encapsulating expert knowledge of the domain. The evaluation of the prototype system has shown that users with a wide range of biological training and experience are able to successfully use the system and interpret and understand the results generated by it. Evaluation showed that the requirements and the interactions with the systems of biologists are substantially different from non-biologists, and preferences for methods of representing the results of argumentation to users varies considerably with even expert biologists differing in the representations they make use of when considering the system output.

Further work will be required to refine the presentation of results and resolve some issues related to levels of detail shown and links to background data required by some users, and to evaluate levels of trust users have in the system.

Acknowledgments

The support of EMAGE and in particular its curator Dr. Jeff Christiansen is greatly appreciated, as is the time given by those who took part in the evaluation. Funding was provided by the EU project Sealife (FP6-2006-IST-027269), and the BBSRC project Argudas (BB/G024162/1).

References

[Atoyan *et al.*, 2006] H. Atoyan, J. Duquet, and J. Robert. Trust in new decision aid systems. In *IHM '06: Proceedings of the 18th International Conference of the Association Fracophone d'Interaction Homme-Machine*, volume 133, pages 115–122, New York, NY, USA, 2006. ACM.

[Baldock and Davidson, 2008] R. Baldock and D. Davidson. *Anatomy Ontologies for Bioinformatics: Principles and Practise*, chapter The Edinburgh Mouse Atlas. Springer Verlag, 2008.

[Bickmore and Green, 2006] T. Bickmore and N. Green. Argumentation for consumers of healthcare. Technical Report SS-06-01, Papers from the AAAI Spring Symposium, AAAI Press, 2006.

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

[Fox *et al.*, 2007] 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.

[Galperin and Cochrane, 2009] M. Y. Galperin and G. R. Cochrane. *Nucleic Acids Research* annual database issue and the nar online molecular biology database collection in 2009. *Nucleic Acids Research*, 37:D1–D4, 2009.

[Gordon, 2007] T. F. Gordon. Visualizing carneades argument graphs. *Law, Probabiliy & Risk*, 6(1-4):109–117, October 2007.

[Jefferys *et al.*, 2006] Benjamin R. Jefferys, Lawrence A. Kelly, Marek J. Sergot, John Fox, and Michael J. E. Sternberg. Capturing expert knowledge with argumentation: a case study in bioinformatics. *Bioinformatics*, 22(8):923–933, 2006.

[Likert, 1932] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

[M^{c}Leod and Burger, 2007] K. M^{c}Leod and A. Burger. Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In N. Guimarães and P. Isaís, editors, *Proceedings of IADIS International Conference Applied Computing*, pages 489–492, Salamanca, Spain, Februrary 2007. IADIS Press.

[M^{c}Leod and Burger, 2008] K. M^{c}Leod and A. Burger. Towards the use of argumentation in bioinformatics: a gene expression case study. *Bioinformatics*, 24:i304–i312, 2008.

[Motro and Rakov, 1998] A. Motro and I. Rakov. Estimating the quality of databases. In *Proceedings of the 3rd international conference on Flexible Query Answering Systems*, pages 298–307, Roskilde, Denmark, 1998. Springer-Verlag.

[Muir, 1987] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man Machine Studies*, 27:527–539, 1987.

[Naumann, 1996] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science, vol. 2261*. Springer-Verlag, 1996.

[Reed and Rowe, 2001] C. A. Reed and G. W. A. Rowe. Araucaria: software for puzzles in argument diagramming and xml. Technical report, Department of Applied Computing, University of Dundee, 2001.

[Shneiderman, 1992] B. Shneiderman. *Designing the user interface: strategies for effective human-computer interaction. 2nd edition*. Addison-Wesley, Reading, MA., 1992.

[Smith *et al.*, 2007] 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:D618–D623, 2007.

[Strausberg, 1999] R. L. Strausberg. *Molecular pathology of early cancer*, chapter The Cancer Geneome Anatomy Project: building a new information and technology platform for cancer research, pages 365–370. IOS Press, 1999.

[Venkataraman *et al.*, 2008] 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:D860–D865, 2008.

[Verheij, 2003] B. Verheij. Dialectical argumentation with argumentation schemes: an approach to legal logic. *Artifical Intelligence and Law*, 11(1-2):167–195, 2003.

[Walton *et al.*, 2008] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, New York, NY, USA, 2008.