

USING ARGUMENTATION TO TACKLE INCONSISTENCY AND INCOMPLETENESS IN ONLINE DISTRIBUTED LIFE SCIENCE RESOURCES

Kenneth M^cLeod
Heriot-Watt University
Edinburgh, Scotland
kcm1@hw.ac.uk

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

ABSTRACT

This paper starts with a discussion of the sometimes contradictory and incomplete nature of distributed online bioinformatics resources. From this we suggest there is a need for a method to evaluate the data contained in such resources before they are used. The method we put forward uses a form of non-monotonic reasoning called argumentation. Arguments are created to help the user weigh up the evidence in favour of and against the biomedical information presented by different Internet-based resources.

KEYWORDS

Bioinformatics

1. INTRODUCTION

The bioinformatics community has access to distributed resources, which are increasing rapidly, both in terms of quantity and size. Regardless of this growth, the data sets in these resources are often necessarily incomplete. There will always exist some time gap between the introduction of new experimental techniques and their exhaustive application, e.g. few gene expression databases cover entire genomes. Inconsistency between the resources is a further problem that must be overcome in order to make effective use of the data available. With these issues in mind, we suggest that bioinformatics provides a suitable domain for the application of non-monotonic reasoning, focusing in particular on argumentation. We propose the use of schemes to model the reasoning of an expert and use that reasoning to allow a system to evaluate the data presented to users. This will allow non-expert users to critically assess the wide array of data before further analysing it. Section 2 discusses the nature of distributed online resources available to life science researchers. Section 3 describes argumentation. Section 4 talks about the work we have done so far, before concluding in Section 5.

2. BIOINFORMATICS

The 2006 annual review of databases in the sub domain of molecular biology (Bateman 2006) showed that 90 new databases were created, and 68 existing databases were significantly changed in 2005. This growth in online distributed databases illustrates the importance of such resources to the life science community. In addition to publishing data, the Internet provides access to tools that analyse that data, thus providing the

mechanism to create new information. For example, GoPubMed (Doms 2005) annotates literature made available through PubMed (www.pubmedcentral.nih.gov) with terms from the Gene Ontology (www.geneontology.org), an ontology published to provide a controlled vocabulary for describing genes and gene products in any organism.

It is likely that the development of the semantic web and grid technologies will further encourage the use of services that use data published through the Internet. Increasing numbers of tools and repositories will have web/grid services alongside the more traditional data sharing techniques like FTP or Perl programmatic interfaces. The web services will be used as part of automated workflows. Projects like myGrid (Goble 2003) have increased the ease of designing and running workflows, and consequently, the frequency with which this is done.

Regardless of whether the online data resources are accessed by an html interface, web service, Perl or FTP, the user must take into account that these resources are not perfect. They feature inconsistent and incomplete information. For example, in the field of gene expression, which discovers the genes active (*expressed*) in the anatomical structures of organisms, 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. If the latest technology is more sensitive than the previous one, it may detect genes in structures where they were not known to exist. At this point, there will be a need to repeat some of the experiments conducted with the old technology, and the gap in knowledge will have temporarily increased.

In addition to incomplete information the researchers must also consider conflicting information. Two research groups may conduct similar experiments, but obtain different results and conclusions. This may be due to experimental error or simply a slight variation in experimental conditions. Despite the variations, both results will be published and entered into (possibly the same) online databases. Therefore the distributed online databases contradict each other, and themselves. For example, in the context of gene expression databases, such as GXD (www.informatics.jax.org), data is collected from various sources. The contributing studies may be based on different experimental setups, for example the use of different probes for the identification of genes. Such variation can lead to conflicting information; a recent version of GXD contains some 1300 instances of particular genes being reported as expressed and not expressed in the same mouse embryo tissue. For example, a GXD query on which genes are not expressed in the mouse brain will report, amongst others, a gene called Tenascin C, since one experiment reported this to be the case. Unless the user also executes a query that asks which genes are expressed in the brain, (s)he will not pick up that there are 14 experiments stored in GXD that report Tenascin C to be expressed in the brain. Clearly, any workflow that is solely based on the first query would be very suspect in its results.

This is not an issue unique to GXD - we simply used it to illustrate the point - but one that needs to be considered when using other resources as well, including the mouse gene expression database EMAGE at the MRC (genex.hgu.mrc.ac.uk). In fact, resources such as GXD and EMAGE that cover the same biological domain (mouse gene expression in this case) provide a good basis for distributed argumentation systems in bioinformatics.

In conclusion, there are a large and growing number of distributed resources available to the life science community, but there is a requirement to evaluate the output from a life science resource before using it, and a need for appropriate help to be given to the user when doing this evaluation.

3. ARGUMENTATION

We believe that an argumentation-based approach (Carbogim 2000) could provide a solution to the problems described in Section 2. Before we describe our proposed solution, let us first briefly discuss argumentation.

An argument (Pollock 2003) is a reason to believe something is true, it is used in dialogue to support or attack a conclusion. Arguments can also attack and defeat each other. Once defeated an argument can be reinstated if the argument that defeats it, is in turn defeated itself.

When presented with the arguments for/against a conclusion, the user can evaluate the evidence and make a decision as to whether or not to believe it. As time passes, new information becomes available, and so new arguments can be created. These new arguments may defeat existing arguments, thus reinstating other arguments. When presented to the user, these changes may alter their perspective, and so alter their opinion

of the conclusion. The reinstatement of arguments is the technique used to handle non-monotonicity, or defeasibility. This is the idea of assuming a conclusion is true until new evidence shows it is not.

In computing terms, argumentation is often used to make computers argue, or to help them assist humans when arguing. By argue, we do not mean an uncontrolled dispute, but instead we are referring to controlled styles of dialogue similar to those used in legal debates. Argumentation has been successfully used in medical informatics, where it was tried as a method of overcoming the weaknesses of traditional mathematical and logical decision-making techniques (Glasspool 2005). Argumentation is used to help a doctor make a decision by providing a list of options and the arguments for and against each option (e.g. Hurt 2003). Often a recommendation is made, but the practitioner makes the final decision. Because it presents results in a manner humans find natural, argumentation has additionally been used to explain decisions to patients that were made by other means (Williams 2006).

The attributes that make argumentation suitable for use in medical informatics also make it applicable to bioinformatics, where it has been employed to help a user evaluate the output of a single resource (Jefferys 2006). Considering the nature of bioinformatics resources, as discussed above, there is a clear intuitive argument to apply this technology not just to a single tool, but also to a range of distributed online tools and databases.

4. ARGUMENTATION: SUGGESTED IDEA AND USAGE

Our proposal is to create a system that can analyse the output of any online resource that it has arguments for. Such a system can work with calls to a single resource, or calls to several resources featured in a workflow. For each query, the system will follow the same reasoning as a human expert in evaluating the resource's output, asking the same questions, and looking at similar metrics. The system will also compare data from different but related resources in order to provide arguments for/against each possible answer to the query.

The process of creating arguments against a particular result would identify the most controversial results, allowing them to be brought to the user's attention. This would be of great assistance when looking for possible errors.

4.1 The creation of arguments

Our arguments capture the reasoning of human experts. The form of documentation we chose was *argument schemes*. On the one hand arguments are presented to users, but few in the life sciences have a background in formal logic or mathematics, so something more natural is required. On the other hand, our arguments have to be used in a real argumentation system, most of which have a logical basis. Thus pure natural language is inappropriate as is an overly theoretic solution. Argument schemes present the obvious compromise.

A scheme is a natural language template for an argument that consists of two parts. The first is an inference rule comprising of a group of premises and a conclusion. The second part is a group of *critical questions* that allow an argument (i.e. instance of the scheme) to be challenged.

Our schemes are based on the Expert Scheme created by (Walton 1997). This scheme models the reasoning that takes place when an expert witness is called in a legal case. The natural language inference suggests that when an expert voices an opinion, we should trust it because they are experts. The critical questions document the most common lines of attack that the opposition would use, for example asking the expert to provide evidence to support their opinion.

For us, the notion of an expert is replaced by the notion of an online resource, such as GXD. We keep the basic presumption that the data held in the resource is accurate, and so we should believe it. Interviewing biologists, and asking them to explain their reasoning processes created our critical questions.

As we produce more schemes, we insert them into a hierarchy. The top of our hierarchy is a scheme applicable to any resource with general questions like: *are you using the latest data set for the resource?* Underneath that we have schemes for groups of resources like gene expression databases. The critical questions are now less general, for example: *Does the gene expression result have an associated image of the result?* Schemes from that level are then specialised to make them appropriate for individual resources in that field, so the EMAGE mouse gene expression database has a scheme featuring the question: *Did the*

EMAGE editor feel confident enough in the result to award it 3 stars? This makes reference to the scoring system that EMAGE editors use to indicate how confident they are in the result. This star system is not used throughout the domain, so an alternative resource like GXD would not have this question in their scheme.

Sometimes resources have more than one scheme associated with them, e.g. the NCBI's BLAST (<http://130.14.29.110/BLAST/index.shtml>). BLAST is an algorithm, implemented in a number of tools, which allow researchers to compare gene/protein sequences to try and find similar genes/proteins, often in different organisms. The results from BLAST do not actually state whether or not BLAST believes proteins to be similar, but instead provides the user with a number of metrics they can use to make the decision. Most researchers will group the results into one of three categories: similar proteins, no similarity between proteins, and proteins that might be similar. One scheme can determine which group each result is entered into. However, some researchers will wish to examine the results in the third group to determine if similarity does exist. A second scheme is required for this.

In addition to extra schemes for a resource, there are links between schemes of different resources. For example, the second BLAST scheme has a question that asks what the output of Pfam (www.sanger.ac.uk/Software/Pfam) shows. Pfam is a separate online biological resource and so has its own scheme.

It is possible to turn the schemes into inference rules and thus use them in most logic-based argumentation systems.

5. CONCLUSION

The online resources available to the life science community contain incomplete, inconsistent and incorrect information. Therefore the data provided by these resources cannot be taken at face value. This is important not just when the results are used directly, but also when they are combined into an automated workflow, where the incorrect data from one resource may mean that all subsequent resource queries are misleading.

We have suggested that one form of non-monotonic reasoning may provide a suitable solution to this problem. For each resource query, argumentation could create arguments for/against every possible result, thus enabling the user to make an informed decision as to which result is correct.

Our initial experiments in this area suggest that this is potentially a hugely beneficial area of research in bioinformatics. It also suggests that the use of so-called argumentation schemes will be the most acceptable form of argumentation to biologists, as it not only provides a reasonable logic-based foundation, but also is easily understood by scientists without a detailed knowledge of logic theory. However, many questions remain to be investigated, such as a more formal assessment of usability and issues of scalability in the context of worldwide resources on the Internet.

ACKNOWLEDGEMENT

Funding by the EU projects Sealife (FP6-2006-IST-027269) and REVERSE (FP6-2006-IST- 506779) is kindly acknowledged.

REFERENCES

- Bateman A., 2006, Editorial, *Nucleic Acids Research*, Vol. 34
- Carbogim D. V. et al, 2000, Argument-based application to knowledge engineering, *Knowledge Engineering Review*, Vol. 15, No. 2, pp. 119-149
- Doms A., Schroeder M., 2005, GoPubMed: Exploring PubMed with the GeneOntology, *Nucleic Acids Research*, Vol. 33, pp. W783-W786
- Goble C. et al, 2003, The myGrid project: services, architecture and demonstrator, *Proceeding of UK e-Science All Hands Meeting*, Nottingham, UK, pp. 959-603
- Glasspool D. W., Fox J., Knowledge, argument and meta-cognition in routine decision-making, *The routines of decision making*, Lawrence Erlbaum, New Jersey, USA, pp. 343-358

- Hurt C. et al, 2003, Computerised advice on drug dosage decisions in childhood leukemia: a method and a safety strategy, *Artificial Intelligence in Medicine*, Protaras, Cyprus, pp. 158-162
- Jefferys B. R. et al, 2006, Capturing expert knowledge with argumentation: a case study in bioinformatics, *Bioinformatics*, Vol. 22, No. 8, pp. 923-933.
- Pollock J., 2003, Defeasible reasoning with variable degrees of justification, *Artificial Intelligence*, Vol. 13 No. 1-2, pp. 233-282
- Walton D., 1997, *Appeal to expert opinion : arguments from authority*, Penn State Press, PA, USA
- Williams M. and Williamson J., 2006, Combining argumentation and bayesian nets for breast cancer prognosis, *Journal of Logic, Language and Information*, Vol. 15, No. 1-2, pp. 155-178