# An analysis and hypothesis generation platform for heterogeneous cancer databases.

Philip Roy QUINLAN [a,1], Alastair THOMPSON [a] and Chris REED [b]
[a] *Dundee Cancer Centre, University of Dundee, Ninewells Hospital, Dundee*
[b] *School of Computing, University of Dundee, Dundee*

**Abstract.** The field of cancer research is now generating vast amounts of data from a variety of high throughput techniques and these have helped to define cancers based on their genetic foundations. As this knowledge on the processes and underlying genetics of cancer improve, these should be factored back into the research and analyses conducted by other researchers. Managing this volume of data, often conflicting, is becoming increasingly challenging for researchers. This work demonstrates an innovative application of argumentation theory within cancer research by providing a framework to accommodate missing data, address critical questions and generate hypotheses. The prototype system has been validated to demonstrate it identifies the same interesting interactions and molecules as researchers, even when certain key data was deliberately withheld from the system.

**Keywords.** application of argumentation, ASPIC+, breast cancer research, automated statistical analysis, hypothesis generation

## 1. Introduction

Argumentation theory has seen many applications, including law [1], medicine [2] and bioinformatics [3,4]. Within this paper we propose another novel application of argumentation in the field of cancer research.

Cancer research generates large volumes of data from an array of different experimental techniques. From this, knowledge of the genetic foundations of cancer have been established allowing the classification of cancers to be improved [5]. Publicly available databases [6] also make it possible for researchers to query similar work conducted by others. With this additional knowledge and access to data come additional requirements when it comes to analysing research data.

This novel application of argumentation aids the researcher by establishing, within an environment of conflict and missing data, what interactions and molecules are interesting and would warrant further investigation. This work has been validated by removing certain data from the system to establish if it would come to the same conclusions as researchers who were in possession of this data.

[1]Corresponding Author: Medical Sciences, Dundee Cancer Centre, University of Dundee, Ninewells Hospital, Dundee, DD1 9SY; E-mail: p.quinlan@dundee.ac.uk

## 2. Background

### *2.1. Cancer Research*

Breast cancer is the most common cancer amongst woman and is increasing in incidence [7] and therefore has been the focus of much research work. One of the aims of this research work is to understand what genes or proteins are associated with larger, more aggressive cancers that do not respond to conventional treatment and ultimately result in poorer survival for the patient. Once the relative expression levels of proteins are obtained, these are usually converted into a boolean value to represent the presence of absence of that protein. For gene data, there are high throughput methods that can allow for the expression of thousands of genes to be quantified from just one sample, providing cancer researchers an invaluable glimpse into the genetic foundations of the cancer. This work [5] has had some success in understanding the differences between breast cancers and has shown breast cancer to be made up of several sub diseases. This improved classification of breast cancer has meant that researchers should conduct their analyses and draw conclusions in the context and knowledge of these sub-types, rather than the more generic breast cancer population.

### *2.2. Tissue Banks*

Tissue Banks provide an invaluable resource to researchers without which much of the progress in cancer research would not be possible [8]. Under ethical and NHS Caldicott Guardian approval, they collect biological samples from patients who are undergoing a clinical procedure and they are asked to donate various tissue samples that are excess to any diagnostic requirements. Clinical data is also collected, such as details on their diagnosis, treatments and other clinical parameters linked to the cancer. This means that over time the tissue bank will contain thousands of cancer samples with full complementary data. As patients are unique and diverse individuals, tissue banks need to collect this associated data so that when a researcher is analysing the data, they can accommodate for differences such as age within their analyses. Tissue banks therefore become more than just tissue repositories, they are also an invaluable data repository, even more so in the current climate of encouraging data sharing. Tissue banks are also well placed to capture any experimental data generated from the tissue samples given to researchers and which over time they can make available to other researchers. By doing this a very comprehensive picture on the makeup of the cancer can be obtained for each patient. The Tayside Tissue Bank (www.tissuebank.dundee.ac.uk), based in the Dundee Cancer Centre has been leading many of the national tissue banking initiatives in this way.

### *2.3. Argumentation*

Argumentation can provide a framework for dealing with scenarios where information is incomplete, inaccurate, often conflicting and where precise modelling may not be possible. Decisions or conclusions are made on the balance of evidence and after examining the for and against arguments. These relationships between attack and support were formalised in an abstract framework [9], which has become the cornerstone of much of the other work in the field. Cancer research is riddled with conflict, uncertainty, ever changing hypotheses and incomplete data and therefore is potentially a very viable candidate

for the application of argumentation. This is exemplified in the use of argumentation in related fields such as bioinformatics [4,3], medical applications or decision support [2,10] and prognosis prediction and treatment[11,12]. A common reason within these applications for the adoption of argumentation is the explanatory power that argumentation offers over more traditional mathematical and AI models. This clarity of why a decision was reached is particularly important when the decision needs to be communicated and understood by the recipient, as was found in the tool for determining if patients should be referred for additional treatment [13]. Within that study, it was found that this explanatory power came at no loss of precision when compared to a mathematical based prediction system [14]. This need for explanation is also extremely important in the field of cancer research, in which the output is often known, (for example, whether the patient lived or died, or whether they had any recurrences of the disease), but the reason it has occurred is unknown. In the field of bioinformatics, argumentation has been used to help guide the researcher through a series of questions to help them draw conclusions [4] and these more recent applications have been based on the Argumentation Service Platform with Integrated Components (ASPIC) [15] framework. A slight alteration to argumentation as the sole method is evident in systems that have used argumentation in combination with more traditional mathematical modelling techniques [11,12]. A similar argumentation framework was designed to ask a series of relevant critical questions from the output of a mathematical model, it was found to significantly improve the overall predictive ability of the model [3]. In relation to this project, argumentation will be utilised to help in two areas. The first scenario is to help bridge missing data and in particular for the presence of a protein when the presence has not been explicitly tested. An argumentation framework can be used to determine if other data that is present, can provide sufficient evidence to conclude the missing data should exist. The second scenario is to then pose critical questions on any subsequent statistical analyses that are performed to ensure any conclusions are logical.

## 3. Motivation

Cancer research is becoming an extremely data rich environment and with tissue banks and data sharing more routine, large repositories of data from the same group of patients have begun to be collected. For knowledge discovery and for the understanding of cancer, this is a fantastic opportunity. However, for a researcher, the complexity in data analyses continues to grow exponentially. With new knowledge and understanding about cancer and the recognition that cancer even from the same organ is not a single disease, there are an increasing number of data points and sub-populations that require consideration, especially if these sub-populations of cancer are believed to behave differently, use different pathways to develop and grow or respond differently to treatments. This highlights the importance of context when analysing the data. For the researcher this increased complexity and number of analyses becomes challenging. A much used but often criticised approach is to use exhaustive data analysis techniques that simply test all data points against each other and in all possible contexts and sub-populations. Although this process will ultimately not miss anything given it tests everything, the statistical validity of the results can be questioned and the researcher is left to shift through potentially thousands of results. The analysis approach should be more selective where the most appropriate tests and hypotheses are tested based on some prior knowledge.

The use of external, publicly available, data could be used as a method to guide data analyses. Even if the researcher could do the analyses required, there is a problem that is hard to overcome. They may be aware themselves of contexts that are important, or have found contexts within the public databases that look interesting, but this data may not be available within the dataset in which the researcher is conducting the analyses. This essentially makes any analysis via any conventional method very challenging. To then further complicate the analysis, as the cancer develops, grows and spreads, different biological processes and responses may be activated. However, the sample that is used to conduct the experiment is usually only taken at one point in time. Therefore the expression levels of genes or proteins can only give the briefest of glimpses into a much longer and larger process.

All of this provides ample justification for the exploration of an alternative analysis process that can handle conflicting or missing information, reason over the processes that are likely to have, or are going to occur, process data from multiple data sources and do so in a way that is transparent to enable the system to explain to the researcher the analyses conducted. Argumentation theory was therefore identified as a suitable framework in which to develop such a system given that it has a proven record within the medical and bioinformatics field at being able to successfully handle similar scenarios to those found within cancer research.

## 4. Architecture

### *4.1. Databases*

Core to the system are several databases that contribute various pieces of data that can be grouped into two categories. The first is held locally within the Tayside Tissue Bank (www.tissuebank.dundee.ac.uk), Aperio (www.aperio.com) and breast pathology database. These provide all the information relating to the samples as well as previously generated research results. This becomes important when the researcher may not necessarily have the data themselves, but it does exist for the same group of samples that are being analysed. The final local database to contain directly relevant data is the breast pathology database. This holds information relating to the patient, such as age, gender, date of birth and importantly all of the data relating to the cancer, such as the grade, tumour size, type and any recurrences.

The other set of data is derived from publicly available online databases. Although this data cannot be used to directly analyse the data generated on the samples, it can be used to advise and guide the analysis process. (1) The pathway interaction database (PID) [6] contains causal data on thousands of interactions between various molecules. Although this data is available via web services, in the first instance, this data will be downloaded in XML format and stored within a local database. This is to ensure the structure is compatible with the required queries and also to ensure stability and control over the dataset during development. (2) Gene expression data, that is available online, that quantifies the expression levels of thousands of genes from cancer patients. This data will be processed into a simplified version and stored in a database to allow for interrogation as required.

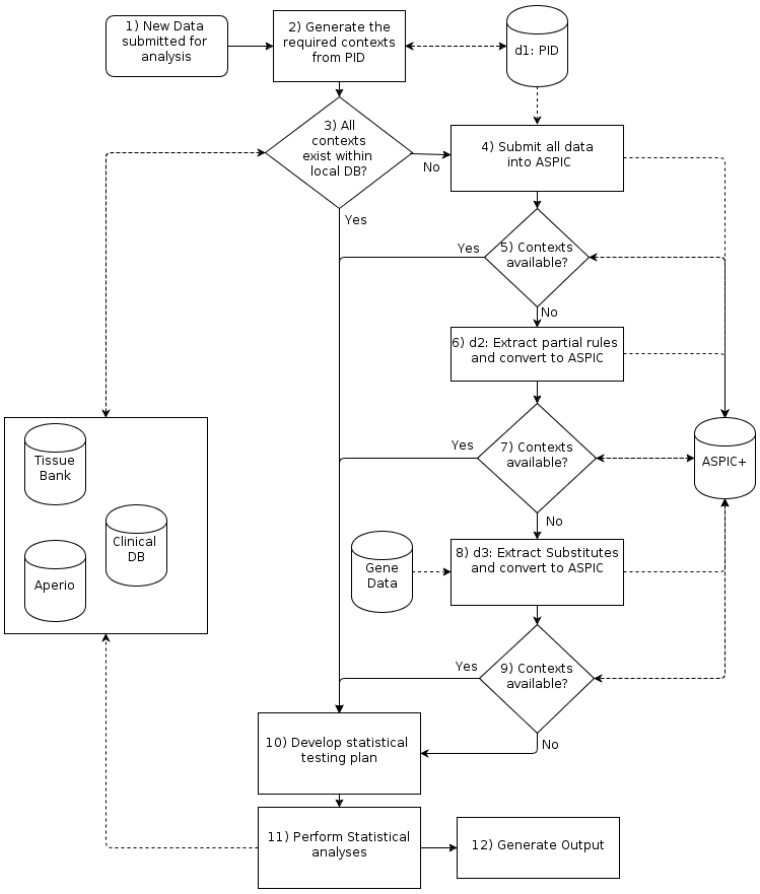


**Figure 1.** Overview of the system components: d1,d2,d3 are database IDs, PID=Pathway Interaction Database

### *4.2. Argumentation Framework*

For the purposes of this work, an extension [16] of the ASPIC [15] framework will be used that has been implemented in Dundee (Mark Snaith 2012 [title] ). The framework will be used in two core functions of the analysis. The first will be to establish if there is sufficient evidence to substitute missing data with that this is available. This is particularly important in the scenario where a context has been found to be important but this data is not within the local dataset. The ability to examine the reasoning behind the decision will enable the system to detect data that could be used as a substitute for the missing data. The second core function of the argumentation framework will be to pose critical questions of the completed analyses to establish the credibility and whether the results are consistent with other projects or previous conclusions. To convert the data from the databases into ASPIC+ compatible rules, there will be an additional module that handles the transitions between the databases and the argumentation framework (Figure 1, Stages 4,6,8).

## 5. Implementation

### *5.1. Databases to ASPIC+*

As rules and premises submitted to ASPIC+ are based on data from external databases, a structured syntax has been generated to ensure the system can transition between both ASPIC+ and the databases. There are two premises, one to represent if a molecule exists and one to represent a data source. A molecule is represented by the ID from the database and prefixed with a 'm' (e.g m1). The data source premises is also represented by the the ID of the database, prefixed with a 'd' (e.g. d1).

Interactions between molecules are managed by detailing the type of interaction by keywords: creates; changes; combines; represents; coexpressed. For example, if the PID has a rule that molecule 1 creates molecule 2, this is represented within ASPIC+ as: m1_creates_m2. This just leaves the situation where a molecule is believed to interfere with an interaction, this introduces the interferes keyword, which is followed by the interaction separated by a colon. So if molecule 3 interferes with molecule 1 creating molecule 2, the syntax would be: m3_interferes:m1_creates_m2.

### *5.2. Data Preprocessing*

The data from the external data sources, such as the pathway interaction database (PID) and the gene array expression data contain a vast amount of static information. Therefore this data is suitable for pre-processing and conversion into a form that will be more readily available at runtime. The gene array expression data is essentially a real value on a scale starting from zero to the tens of thousands. This makes it hard to make any direct comparisons and so they need to be grouped in some way. The method of using quartiles was chosen as the most suitable method for the needs of this project. This makes it possible to ask when patients have gene A expression in the highest quartile, which other genes are also within the highest quartile. For each gene the values that determine the boundaries of the quartiles are established. Subsequently for each sample and for that gene, the expression value is taken and converted to a quartile based on these boundaries.

The other external data is from the PID that contains causal relationships between molecules. As such, this data will be the core data from which rules will be generated and entered into ASPIC+. Argumentation schemes have been generated based on the syntax in 5.1, to represent the causal rules from PID: Creation (Figure 3); Alteration (Figure 4); Combination (Figure 5), alongside each scheme are the associated rules that would be created based on the syntax described. This data along with the ASPIC+ formatted rules are stored within a local database, so they can be used to prime ASPIC+ before an analysis run. A simplified example will be used to demonstrate the various stages, using a small subset of rules that were extracted from the PID.

### *5.3. New data submitted and calculating the contexts*

All the processing that has occurred to date has been prior to any new analysis process being initiated. To demonstrate the use of the system a simple example will be used to go through all the processes involved in Figure 1. For the purpose of this example, the data to be analysed is a protein called HSPB1 (m101395). As described in Section 3, it

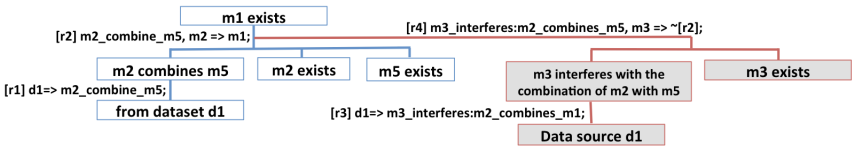


**Figure 2.** Argumentation scheme for the creation of molecule m1 by molecule m2

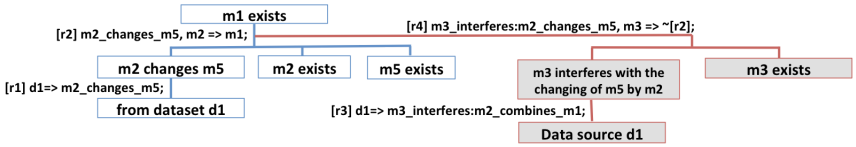


**Figure 3.** Argumentation scheme for the alteration of molecule m5 to molecule m1

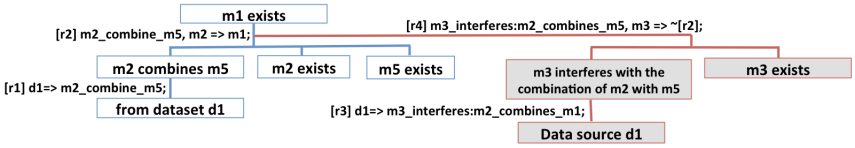


**Figure 4.** Argumentation scheme for the binding of molecule m2 with molecule m5

**Table 1.** Rules generated from the argumentation schemes

| Rules | Premises |
|---|---|
| [r1] d1 =>m101456_creates_m101395 | m101395 |
| [r2] m101456_creates_m101395, m101456 =>m101395 | m202413 |
| [r3] d1 =>m101398_changes_m101395 | d1 |
| [r4] m101398_changes_m101395, m101398 =>m101395 | |

is important to understand the contexts when analysing data and for HSPB1 these are obtained from the PID (Figure 1, Stage 2). The contexts are derived from two interaction types from within the PID, those that have been responsible for the either the creation or alteration of protein HSPB1. In this example, m101456 (Estrogen/Estrogen/ER alpha/ER alpha) creates HSPB1 and m101398 (MAPKAPK2) alters HSPB1 and these are therefore identified as the contexts required for the analyses. Now that these have been identified it is required to check if this data is available (Figure 1, Stage 3) by querying for their existence within either the Tayside Tissue Bank or Aperio databases. In the unlikely event that all the data is available, the process can move straight to statistical testing (stage 10). However, in this example, this data is not present for the same group of samples for which we have the HSPB1 data. The only other protein data available, is ER alpha (m100868) and IGF1R (m201886).

This then becomes the first point in which argumentation can be utilised to aid the process. As it stands, two contexts have been identified, but as the data does not exist,

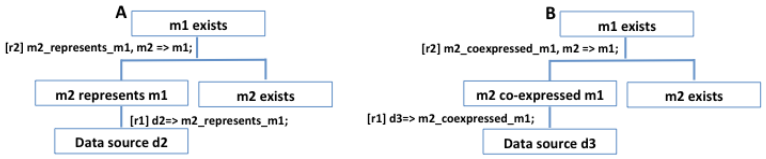


**Figure 5.** Argumentation scheme for the representation of a valid substitution from A) partial data B) gene expression data

traditional analysis packages would not be able to continue any further testing. ASPIC+ will now be used to to establish if there is any evidence for either ER alpha (m100868) or IGF1R (m201886) to be used as substitutes for m101456 (Estrogen/Estrogen/ER alpha/ER alpha) or m101398 (MAPKAPK2) within the analyses. Table 1 summarises the rules and the premises that will be used. Two independent queries will be made within ASPIC+ (Figure 1, Stage 4), for m101456 and m101398. Given the current set of premises and rules, ASPIC+ cannot conclude that m101456 (Estrogen/Estrogen/ER alpha/ER alpha) or m101398(MAPKAPK2) can be substituted by the available data. The only way for the analysis to progress will be if other rules or premises can be generated and this is explored in the following stage (Figure 1, Stage 6).

### *5.4. Generating rules from partial data*

The data held in PID contains the interaction details of thousands of proteins. Two types of these interactions can make finding a direct match between local data and that found in the PID difficult. Combination interactions are when two individual molecules, such as proteins, bind together to form a new molecule. Therefore to obtain a direct match, local data must exist for both of the sub-components. In the HSPB1 example m101456 (Estrogen/Estrogen/ER alpha/ER alpha), is an example of this combination, with the molecule constituting of two sub-components, estrogen and ER Alpha. To date, it has not been possible to use any of the existing data as a substitute for m101456 (Estrogen/Estrogen/ER alpha/ER alpha), even though one of the sub-components, ER alpha, does exist.

In this scenario, the system attempts to find if a sub-component can be a substitute for the presence of the combined molecule. There are two situations in which this is acceptable. The first is when the components that form the combined molecule are always found together and there is no evidence for anything preventing the combination from occurring. Unfortunately this rule is not satisfied, as ER alpha and estrogen are found to be within complex molecules where either are not present.

The second scenario in which ER Alpha could be a substitute for the combined molecule is when the combined molecule contains molecules which have been defined as compounds within PID. Although the compound components can be important, the data generated and used for analysis will almost exclusively be protein data and therefore making a direct match on the combined molecules containing compound components extremely rare. It is only once the other avenues have been explored for finding a match that the system removes any compound components. Once the compound components have been removed from m101456 (Estrogen/Estrogen/ER alpha/ER alpha) the system can conclude that ER alpha alone can be a substitute for the combined molecule m101456 (Estrogen/Estrogen/ER alpha/ER alpha).

**Table 2.** Output from calculating co-expression with IGF1R

| MoleculeID | Expected | Measured | Percentage Swing |
|---|---|---|---|
| m101398 | 87.75 | 102 | 12% |
| m100831 | 87.75 | 94 | 5% |
| m203224 | 87.75 | 107 | 16% |

This is formalised by using the scheme in Figure 5A, which allows the reasoning to be captured with ASPIC+. This is important when presenting the data to the researcher as they can interrogate the reasoning behind any of the substitutions. This rule along with any others that have been generated during this process, are added to the ASPIC+ ruleset and ASPIC+ is queried for evidence that m101456 (Estrogen/Estrogen/ER alpha/ER alpha) or m101398(MAPKAPK2) can be substituted by the available data. Unsurprisingly, given the inclusion of the rule based on Figure 5A, it can now be concluded that m101456 (Estrogen/Estrogen/ER alpha/ER alpha) can be substituted by m100868 (ER Alpha). There is still however no evidence to support a substitution for m101398(MAPKAPK2). There is one last avenue the system will explore to look for alternatives and this is captured with Figure 1 (Stage 8).

### *5.5. Generating rules from the Gene Array expression data*

The gene array expression data has so far been processed into quartiles but has yet to be used in any of the analyses. The data is utilised in the scenario where it has been established that important context data still cannot be substituted by available data, even once accounting for partial data. Therefore this represents the last chance to find substitutes for the missing contexts (Figure1, Stage 8).

The gene array data contains vast amounts of data on the expression levels of thousands of genes. Given this, it is possible to ask questions such as, when the gene IGF1R is highly expressed, what other genes are also highly expressed. The ability to ask this question in this scenario is important as it can highlight genes that may be suitable as substitutes. This process is always left as the last result because as previously mentioned, the data most likely to be analysed is protein levels and this data is the expression levels of genes. The reason this data can be useful, is because gene expression is the first in a multi-step process for the creation of proteins [17] . Therefore any suggested substitutions are valid but come with a lower confidence, therefore this stage of the system is only reached if the other methods have failed to find substitutes.

To find the substitutes, for each of the data held, apart from HSPB1, the gene array expression database is queried for genes that are expressed at similar levels. The data has been transformed into quartiles, therefore the system will look for similarities within the quartile groups. Using IGF1R as an example, the database is queried for substitute genes that are expressed within quartiles 2-4 at the same time as IGF1R. If there was no association between the genes, it would be expected that a substitute gene would have an even representation of the quartile groups within the IGF1R 2-4 quartile group. In the example, IGF1R has 117 cases that fall within quartiles 2-4. If everything were distributed evenly, geneA should have 75% or 88 cases within quartiles 2-4. If there are more than 88 cases within quartiles 2-4 for Gene A, it would indicate some correlation between the expression levels of gene A and IGF1R. It was decided that for the difference to be considered significant, there should be a 10% increase. As an example, Table 2 shows the

**Table 3.** Total input into ASPIC+ for processing

| Rules | Premises |
|---|---|
| [r1] d1 =>m101456_creates_m101395 | m101395 |
| [r2] m101456_creates_m101395, m101456 =>m101395 | m100868 |
| [r3] d1 =>m101398_changes_m101395 | m201886 |
| [r4] m101398_changes_m101395, m101398 =>m101395 | d1 |
| [r5] d3 =>m100868_represents_m101456 | d2 |
| [r6] m100868_represents_m101456, m100868 =>m101456 | d3 |
| [r7] d2 =>m201886_coexpressed_m101398 | |
| [r8] d2 =>m201886_coexpressed_m203224 | |
| [r9] m201886_coexpressed_m101398, m201886 =>m101398 | |
| [r10] m201886_coexpressed_m203224, m201886 =>m203224 | |

output from the process, detailing the expected number versus those actually within quartiles 2-4 and the percentage swing. From this, it can be concluded that IGF1R can also be used as a substitute for m101398 and m203224. Therefore a new rule can now be entered into ASPIC+, using the argumentation scheme detailed in Figure 5B. This combined with all the other previous processing steps, involves the rules and premises set out in Table 3 now being present with the ASPIC+ framework and the query for the contexts can be re-run a final time (Figure 1, Stage 9). Now all the contexts can be substituted by data available, by using partial data to represent m101456 (Estrogen/Estrogen/ER alpha/ER alpha) and co-expressed data to represent m101398 (MAPKAPK2). As the system has now established the contexts and any substitutes, the system moves on to creating and running the statistical testing plan.

### *5.6. Developing the statistical testing plan and running the tests*

Now that the contexts and substitutes have been established, the process of determining what tests should be conducted (Figure 1, Stage 10) and what data to use is relatively straightforward and based on previously implemented work [18]. The main difference here, however, is that the exhaustive and repetitive nature of the testing has been removed and the previous steps have formed very specific tests to run. The focus of this system is on patient survival after diagnosis as an end point which limits the statistical tests, to test for significance the Log Rank test is used and the Kaplan Meier method used to graphically represent the survival data. Given these restrictions the number of statistical tests to be run will be small, however some demanding questions are asked, Test 1 asks whether HSPB1 results in any change in survival when created by ER Alpha. Test 2, asks whether HSPB1 that has been created by ER Alpha and then subsequently altered by MAPKAPK2 results in a change in survival.

As with any statistical tests, a cut off must be decided to indicate significant results within the system. A p value below 0.1 as is marked as 'worthy of further analysis', below 0.05 as 'significant' and below 0.01 as 'highly significant'. Test 2 produces a p value of 0.08 and therefore within the 'worthy of further analysis' group. As such, further tests are conducted against other standard predictors of survival, such as tumour size,

tumour grade and whether the tumour has spread to the lymph nodes. The system uses the results of these to ask some critical questions of the previous results and therefore establishing the strength of the previous significance.

## 6. Evaluation and Future Work

This new method of testing clinical research data can bring several benefits. The questions that were asked and statistically tested were very specific, limited in numbers and intelligently based on previous knowledge from publicly available databases into the the analysis process, providing valuable context to the researcher. The real distinguishing feature is the ability to handle missing or incomplete data. This allows the researcher to make informed decisions on whether the results bear any real significance or importance.

The data that was used for this example came from a real example in which researchers had been struggling to understand the results. The protein HSPB1 was being examined and analysed using an exhaustive data analyses program [18]. It found numerous associations, within numerous sub-populations and combinations of these. Data was also available on the altered version of HSPB1, and the same tests as the normal form of HSPB1 were conducted, further adding to the complexity. This left the research team with a large amount of work to establish whether these significant results made any sense, a process that spanned many months. It is from this work that the example in this paper was generated. However, the system was deliberately given a handicap, the data on the altered version of HSPB1 was withheld. The intention was to see if the system could replicate the conclusions made by the research team, even when it had less data available than the research team.

The system found that the basic form of HSPB1 was not significant (p=0.56) in the context of ER Alpha. However, it did find that this conclusion changed when looking at the altered form of HSPB1 within the context of ER Alpha, as it was marked as 'worthy of further investigation' (p=0.08). This was the same conclusion as the research team who were in possession of the complete dataset. This clearly demonstrates the potential power of the system to deliver clear and concise information to researchers, to either confirm or generate new hypotheses, in an environment of missing and incomplete data.

Further work is required to refine the processes and to complete the system to a fully operational state. In particular the gene expression mining section may require further testing to check the validity of rules generated. Additions will also be included to increase and improve the critical questions asked after the completion of the analysis. The system will be tested in other similar scenarios where the conclusions made by the system can be compared to those made by research teams before it is fully utilised on previously untested datasets in which the conclusions are yet to be made.

## 7. Acknowledgements

The authors would like to acknowledge, Dr Lee Jordan and Dr Colin Purdie for their help with pathology data, Dr Phil Coates for assistance with scientific data, Mr Mark Snaith for the implementation of ASPIC+ and Breast Cancer Campaign.

## References


[1] Trevor Bench-Capon and Henry Prakken. Using argument schemes for hypothetical reasoning in law. *ARTIFICIAL INTELLIGENCE AND LAW*, 18(2):153–174, 2010.

[2] John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. Argumentation-based inference and decision making—a medical perspective. *Intelligent Systems, IEEE*, 22(6):34–41, 2007.

[3] Benjamin R. Jefferys, Lawrence A. Kelley, Marek J. Sergot, John Fox, and Michael J. E. Sternberg. Capturing expert knowledge with argumentation: a case study in bioinformatics. *BIOINFORMATICS*, 22(8):924–933, 2006.

[4] Kenneth McLeod and Albert Burger. Towards the use of argumentation in bioinformatics: a gene expression case study. *BIOINFORMATICS*, 24:i304–i312, 2008.

[5] Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, 378(9805):1812–1823, November 2011.

[6] Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H. Buetow. Pid: the pathway interaction database. *Nucleic Acids Research*, 37:674–679, 2009.

[7] http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/uk-breast-cancer-incidence-statistics, 2011.

[8] Alastair Thompson, Keith Brennan, Angela Cox, Julia Gee, Diana Harcourt, Adrian Harris, Michelle Harvie, Ingunn Holen, Anthony Howell, Robert Nicholson, Michael Steel, and Charles Streuli. Evaluation of the current knowledge limitations in breast cancer research: a gap analysis. *Breast Cancer Research*, 2008.

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

[10] David W. Glasspool, John Fox, Ayelet Oettinger, and James Smith-Spark. Argumentation in decision support for medical care planning for patients and clinicians. In *AAAI Spring Symposium Series*, 2006.

[11] Matt Williams and Jon Williamson. Combining argumentation and bayesian nets for breast cancer prognosis. *Journal of Logic, Language and Information*, 15:155 – 178, 2006.

[12] Matt Williams and Anthony Hunter. Harnessing ontologies for argument-based decision-making in breast cancer. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 254–261. IEEE Computer Society, 2007.

[13] 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. In *Methods of Information in Medicine*, volume 315-322, 2001.

[14] Jon Emery, Robert Walton, Michael Murphy, Joan Austoker, Pat Yudkin, Cyril Chapman, Andrew Coulson, David Glasspool, and John Fox. Computer support for interpreting family histories of breast and ovarian cancer in primary care: Computer support for interpreting family histories of breast and ovarian cancer in primary care: comparative study with simulated cases. *BMJ*, 321:28–32, July 2000.

[15] Leila Amgoud, Lianne Bodenstaff, Martin Caminada, Peter McBurney, Simon Parsons, Henry Prakken, Jelle van Veenen, and Gerard Vreeswijk. Deliverable d2.6 - final review and report on formal argumentation system. Technical report, 2006.

[16] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, pages 93–124, 2010.

[17] Francis Crick. Central dogma of molecular biology. *Nature*, 227, 1970.

[18] Philip R Quinlan, Alastair Thompson, and Chris Reed. Inspire: An integrated agent based system for hypothesis generation within cancer datasets. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 587–590, 2008.