Measuring the Implications of the D-basis in Analysis of Data in Biomedical Studies

Kira Adaricheva[1,2,⋆], J.B.Nation[3,⋆,⋆⋆], Gordon Okimoto[4], Vyacheslav Adarichev[1,⋆⋆],
Adina Amanbekkyzy[1], Shuchismita Sarkar[1], Alibek Sailanbayev[1],
Nazar Seidalin[5,⋆⋆], and Kenneth Alibek[6,⋆⋆]

[1] Nazarbayev University, School of Science and Technology, Kazakhstan
[2] Yeshiva University, New York, USA
[3] University of Hawaii, USA
[4] University of Hawaii Cancer Center, USA
[5] Medical Holding, Astana, Kazakhstan
[6] Nazarbayev University, Graduate School of Medicine

Abstract. We introduce the parameter of relevance of an attribute of a binary table to another attribute of the same table, computed with respect to an implicational basis of a closure system associated with the table. This enables a ranking of all attributes, by relevance parameter to the same fixed attribute, and, as a consequence, reveals the implications of the basis most relevant to this attribute. As an application of this new metric, we test the algorithm for D-basis extraction presented in Adaricheva and Nation [1] on biomedical data related to the survival groups of patients with particular types of cancer. Each test case requires a specialized approach in converting the real-valued data into binary data and careful analysis of the transformed data in a multi-disciplinary environment of cross-field collaboration.

Keywords: Binary table, Galois lattice, implicational basis, D-basis, support, relevance, gene expression, survival, response to treatment, immune markers, blood biochemistry, infection.

Knowledge retrieval from large data sets is an essential problem in economy, biology and medical sciences. The data is often recorded in tables with rows consisting of the objects and columns of the attributes. The dependencies existing between subsets of the attributes in the form of *association rules* can uncover the laws, causalities and trends hidden in the data.

In data mining, the retrieval and sorting of association rules is a research problem of considerable interest. The benchmark algorithms, such as *Apriori* in Agrawal et al. [5], have the complexity that is exponential in the size of a table. Moreover, the number of association rules is staggering, and thus it requires further tools for filtering to obtain a short subset of rules that are significant. There are no strong mathematical results confirming a particular choice of such short subsets, and numerous approaches to the filtering process are described in various publications devoted to the topic. See, for example, Kryszkiewicz [14] and Balcázar [7].

One particular subset of association rules, the *implications*, or rules of full confidence, merit particular attention in data mining, as well as being the center of on-going theoretical study, supported by a number of strong mathematical statements. This could be explained by the fact that implications constitute one of the facets of closure systems. In particular, they closely relate to the structure of finite lattices.

Representation of a binary table and its concept (Galois) lattice *via sets of implications* continues to be a primary research goal of concept analysis (FCA). The target

[⋆] Partial support provided by grant N 13/42 (2013–2015) of Nazarbayev University.

[⋆⋆] Partial support provided by grant N 0112PK02175 (2012–2014) of Ministry of Healthcare and Social Development of RK.

for many years was the retrieval of the *canonical basis*, or Guigues-Duquenne basis, of implications for closure systems defined by a Galois connection on the binary table. Nevertheless, recent results confirm that algorithmic solutions to such a task have complexity that is at least exponential in the size of the table; see, for example, Distel and Sertkaya [10] and Babin and Kuznetsov [6].

In Adaricheva and Nation [1], the authors suggest using a new type of basis, called the *D-basis*, which was introduced in Adaricheva, Nation, and Rand [2]. This basis has a lattice-theoretical flavor, for its generating notion is that of a minimal cover of a join irreducible element in a finite lattice. The D-basis is usually a proper subset of the *canonical direct unit basis* (this latter is different from the Guigues-Duquenne basis, see Bertet and Monjardet [8]), while it enjoys the property of being *ordered direct*.

The advantage of this basis in relation to the canonical basis, for the representation of a binary table, is in the possibility of reducing the task to dualization of an associated hypergraph. It is known that the hypergraph dualization problem has a sub-exponential algorithmic solution, see Fredman and Khachiyan [11]. The algorithm in [1] avoids generating the Galois lattice from the table, and only uses the arrow relations, which can be computed in polynomial time, to produce a hypergraph for each requested attribute. In that way, the existing code for hypergraph dualization, such as in Murakami and Uno [15], can be borrowed for execution.

In the current paper, we employ the code implementation of this algorithm as a working approach for data analysis in biomedical studies.

Since 2013 we have been working with two data sources, both connected with cancer research. One of them is the data sets provided by the bio-informatics group at the University of Hawaii Cancer Center, which relate the gene expression of patients with various types of cancer with their survival parameters. Another source is provided by medical research group at Medical Holding in Astana, Kazakhstan. The data relates the immune, viral and blood parameters of patients with brain tumors with their response to a new regimen of treatment.

While both data sets are essentially different sorts of real-valued medical data, we developed customized approaches to convert them into binary tables. In the case of the Astana research group, we dealt with temporal data, which included several measurements of parameters during the time of treatment of patients. The target of our tests was to reveal possible connections between the dynamics of sets of several parameters with the response to treatment. In analyzing the data from Honolulu, we worked in close collaboration with the Hawaii bio-informatics group, applying the implicational algorithms after the genetic data had been reduced to manageable size by other methods.

As in data mining, the main obstacle to analysis of binary tables *via* their representation by implications is the impressive number of implications in the basis. The algorithm in [1] allows us to retrieve only those implications $X \to b$ in the D-basis that have a fixed attribute b as a conclusion. In our case, b would represent a particular indicator, for example that a patient belongs to the group of long-survivors, say those who lived for longer than 1300 days beyond the day of diagnosis. Such subsets of the basis may contain close to 1500 implications when the table has just twenty attributes (columns). This number can easily increase to 1,000,000 when the number of columns is in the range of 250.

Our goal was to identify small groups of parameters whose appearance in the requested sector of the basis indicated the influence of such groups to the target parameter

b. In order to rank the implications from the subset of the basis having b as a conclusion, we introduced new metric for the attributes of the table. We call the new metric the *relevance* of attribute a with respect to b, and it is computed based on frequency of a appearing in the antecedents of implications related to b in two bases: one for original table, and the other for the table, where attribute b is replaced with its complement $\neg b$. The computation of this parameter also takes into account the support of each individual implication in the basis where a appears.

After computing the relevance parameter for all attributes of the table, one can rank the implications $X \to b$ by taking the average of relevance parameters for all $x \in X$. For each individual data set, it is up to a specialist in the data to establish the lower threshold for the relevance parameter of implications, to separate the small portion of them which might have impact for further study.

We believe that our testing provides some first insights into the possibilities for using implication bases in biomedical studies that involve relatively large data sets.

One of the main achievements of our tests was to demonstrate that the algorithm can handle tables with the number of columns/attributes exceeding those reported in the literature, with respect to canonical or canonical direct unit bases; see for example Ryssel, Distel and Borchmann [16]. We had successful runs of the algorithm on a table with 287 columns, for medical data with relatively high density (proportion of ones in the table), while for less dense data sets, such as transaction tables, there were successful runs for tables with more than 500 attributes.

The paper is organized as follows. We provide the background information on closure operators and associated implicational bases, as well as their connection to binary tables, in the first section. Discussion of the D-basis and related information on other bases used in applications is given in Sec. 2. In the third section we introduce the definition of the parameter of the relevance of an attribute with respect to the fixed target attribute of the binary table. The computation of this parameter is illustrated in Sec. 4, which uses as input the gene expression data related to ovarian cancer provided by the University of Hawaii Cancer Center. A larger data set of the same type is discussed in the next section, where we also discuss some variations in computation of relevance metric. In Sec. 6 we discuss the test results on the temporal medical data provided by Medical Holding in Astana, related to a group of patients with brain tumors. The final section gives an overview of future testing and collaboration.

1 Short introduction to implications

A *closure operator* ϕ on a set A is an increasing, monotone and idempotent function $\phi : 2^A \to 2^A$. It is well known that any closure operator ϕ defined on finite set A can be fully represented by a *set of implications* $X \to y$ with $X \subseteq A$, $y \in A$. Any individual implication $X \to y$ can be considered as partial information about ϕ, saying that the ϕ-closure of X, i.e., the set $\phi(X) \subseteq A$, contains y. Implications $X \to y$ are also called *unit* implications, indicating that a single element y is on the right side of the arrow symbol. The unit implications can be *aggregated*: if there are several unit implications with the same left side X, then one can take the union of the right sides into a subset Y and represent these unit implications *via* $X \to Y$. However, for the algorithms used in this paper, it is better *not* to aggregate the unit implications.

The standard approach for the study and storage of the data related to a closure operator is to record an essential subset of the set of all implications of ϕ, called a *basis*, from which all valid implications (and thus the closure operator itself) can be recovered. There are many types of bases which have been targets for theoretical research, such as the canonical basis of Guigues-Duquenne [12] or the canonical direct unit basis; see the survey article [8].

Another type of basis, called the *D-basis*, was introduced in [2]. This basis is a subset of the canonical direct unit basis, and tends to be noticeably shorter. In our tests, the size of the D-basis (the number of implications) was on the average about 30% shorter than the size of the canonical direct unit basis. On the other hand, the canonical direct unit basis is *direct*, meaning that closures of sets can be computed in one pass. The D-basis retains this property in a slightly modified form, called *ordered directness.*

One special case of the closure operator exists in any data presented by a binary table. By a binary table we understand a triple (U, A, R), where $R \subseteq U \times A$ is a relation between sets U and A, where U is the set of objects (corresponding to rows of the table) and A is the set of attributes (corresponding to columns). If $r = (u, a) \in R$, then the position in row u and column a is marked by 1. This can be interpreted as object u possesses attribute a. Otherwise, the position is marked with a 0.

In order to recover the closure operator on the set of attributes, defined by a given binary table, we introduce two functions between subsets of attributes and objects.

The *support function* $S_A : 2^A \to 2^U$ is defined, for every $X \subseteq A$, by $S_A(X) = \{u \in U : (u, x) \in R \text{ for all } x \in X\}$. Thus, row u is in the support of set of columns X, if all intersections with columns from X, along this row, are marked by 1, or equivalently, if the object u possesses all the attributes from X.

Similarly, the support function $S_U : 2^U \to 2^A$ is defined for all $Y \subseteq U$ as $S_U(Y) = \{a \in A : (y, a) \in R \text{ for all } y \in Y\}$.

It is straightforward to show that the operator $\phi_A : 2^A \to 2^A$ defined as $\phi_A(X) = S_U(S_A(X))$ for $X \in 2^A$ is, in fact, a closure operator on A. Any implication $X \to y$ which holds ϕ_A can be directly interpreted from the table as follows: for each row of the matrix, whenever all intersections of this row with columns from set X are all marked by 1, the position at column y is also marked by 1. Note that the actual number of rows where intersections with X are marked by 1 is usually just a portion of total number of rows, and the set of such rows will be denoted $\sup(X)$, instead of $S_A(X)$, to match the notation used in data mining literature.

Let us illustrate these concepts in the following example. Consider the table with a set U of 6 objects and a set A of 7 attributes.

Table 1.

	b	a_1	a_2	c_1	c_2	y	z
1	0	1	0	1	0	0	0
2	1	0	0	0	0	1	0
3	0	0	1	0	1	0	0
4	0	0	0	1	1	0	0
5	0	0	0	1	1	0	0
6	1	1	0	0	1	1	0

Consider $X = \{a_1, c_2\} \subseteq A$. Then $S_A(X) = \sup(X) = \{6\}$ and $\phi_A(X) = S_U(S_A(X)) = \{a_1, c_2, b, y\}$. Hence, we will have implications $X \to b$ and $X \to y$ in the set of all implications describing the operator ϕ_A. The logical statement "*if* X *then* y" holds in all rows of the matrix, while assumption X holds only in row 6.

From the point of view of data mining, the implication $X \to y$ is an *association rule* between columns of the given matrix with the support parameter $\frac{\sup(X \cup y)}{|U|} = 0.17$, which is just a normalized version of the support, showing the relative frequency of the rows where all attributes from X are marked.

The second essential parameter used for measuring the association rules is the *confidence*:

$$c(X \to y) = \frac{\sup(X \cup y)}{\sup(X)}.$$

If $X \to y$ is an implication, then the confidence is always 1, i.e., the highest among all possible values of this parameter. In general, an association rule may have confidence strictly lower than 1, and a lower bound threshold is used to filter the association rules of importance. For example, we may consider the association rule $c_1 \to c_2$, for which the normalized support is $\frac{1}{3}$, and the confidence is $\frac{2}{3} = 0.66$. This association rule might be discarded from consideration, assuming that the lower bound threshold is established, say, at $c = 0.75$. Among all association rules which can be considered for the attributes of the tabled data, the implications can be characterized as those with the confidence of 1.

Having established the connection with the field of data mining, we will deal in the sequel only with the implications describing the closure operator ϕ_A on the set of attributes of a binary table.

2 Comparison of the D-basis with other bases for the purposes of table description

The algorithm in [1] enables us to obtain the D-basis for the set of implications defining the operator ϕ_A on the set of attributes of a binary table. A critical difference with the other existing algorithms is that, instead of creating an intermediate algebraic structure known as a concept (Galois) lattice, or equivalently, finding all ϕ_A-closed sets, this algorithm retrieves only partial information about the structure in the form of *up-arrows, down-arrows* and *up-down-arrows*, which replace some of the 0-entries of the table. This additional information is enough to form an instance of the well-known *hypergraph dualization* problem, for which algorithmic solutions already exist and are realized in fast-executed computer programs; see, for example, Boros et al. [9] and Murakami and Uno [15].

Another essential difference between the structure of the D-basis and, say, canonical basis of Guigues-Duquenne, is that the D-basis is oriented toward finding, for any fixed $b \in B$, all subsets $X \subseteq A$ such that $X \to b$ is an implication for the operator ϕ_A, and X satisfies some irreducibility property with respect to b. In contrast, finding the canonical basis requires finding all *pseudo-closed* sets of operator ϕ_A, which will serve as antecedents of implications $X \to y$, and this search is irrelevant of what we want to find as the right side of the implications.

One consequence of the irreducibility property for $X \to y$ in the D-basis is that $X' \to b$ is not longer an implication for ϕ_A, for any proper subset $X' \subset X$. The latter property also holds for implications included into the canonical direct unit basis mentioned earlier. At the same time, the irreducibility property required for implications of the D-basis is stronger, which explains why the D-basis is normally a proper subset of the canonical direct unit basis.

Recently, U. Russel et al. [16] proposed a method of retrieval of the canonical direct unit basis that would employ the hypergraph dualization algorithm. It does not employ the D-relation, which we use for the purposes of obtaining the D-basis. The D-relation is a binary relation that can be computed in polynomial time in the size of the table, using the information about the up- and down-arrows mentioned above. This allows us to reduce the size of the hypergraphs for which the dualization should be computed, compared to the algorithm in [16].

3 Measurement of relevance of implications with respect to a fixed attribute

In this section we describe our new approach to the measurement of the implications in a particular implicational basis, with the goal of distinguishing a small subset of implications relevant to a sector of the basis targeting a particular fixed parameter b in the set of attributes.

Given any closure system (A, ϕ) on the set A, and any *unit* implicational basis β defining this closure system, we can define a subset $\beta(b) \subseteq \beta$ with respect to any element $b \in A$ as follows:

$$\beta(b) = \{(X \to y) \in \beta : \phi(y) = \phi(b)\}.$$

For example, when β is the D-basis of the operator ϕ_A for the table given in Sec. 1, we have $\beta(b) = \{b \to y, y \to b\} \cup \{\{a_1, c_2\} \to t, \{a_2, c_1\} \to t : t = b, y\}$. Formally, column y can be stripped from the table, since it is identical to column b, which implies $\phi_A(y) = \phi_A(b)$. One can find the D-basis on the table without y, then extend the information we know for column b to its twin column y.

The algorithm presented in [1] is based on the retrieval of $\beta(b)$, for each $b \in A$, where the closure system is defined on the set of attributes A of a given binary table, and where β is a D-basis of this closure system. It is critical that the retrieval of the basis is done separately for each attribute $b \in A$, so that parallel processing could be done to optimize the required time to obtain the whole basis.

More often, though, the whole basis is not what is needed in a particular study, and the implications of the form $X \to b$, for some particular fixed $b \in A$, are of higher importance than others. Then the choice of the basis will be based on the possibility to compute $\beta(b)$ much faster than the whole β.

When $\beta(b)$ is available, the main task is to rank the implications thus obtained with respect to relevance of antecedent X to attribute b.

From the extensive list of parameters known for the filtering the association rules in data mining, the only parameter that can be applied for ranking of implications is the parameter of *support*. Indeed, while many parameters in data mining make extensive use of the parameter of confidence, that does not apply in the case of implications, as observed in Sec. 1.

We believe that, for each attribute $a \in A \setminus b$, the important parameter of relevance of this attribute to $b \in A$ is a parameter of *total support*, computed with respect to basis β:

$$\mathrm{tsup}_b(a) = \Sigma\{\frac{|sup(X)|}{|X|} : a \in X, (X \to b) \in \beta\}.$$

Thus $\operatorname{tsup}_b(a)$ shows the frequency of parameter a appearing together with some other attributes in implications $X \to b$ of the basis β. The contribution of each implication $X \to b$, where $a \in X$, into the computation of total support of a is higher when the support of X is higher, i.e., column a is marked by 1 in more rows of the table, together with other attributes from X, but also when X has fewer other attributes besides a.

While the frequent appearance of a particular attribute a in implications $X \to b$ might indicate the relevance of a to b, the same attribute may appear in implications $X \to \neg b$. The attribute $\neg b$ may not be present in the table and can be obtained by converting the column of attribute b into its complement.

Let $\beta(\neg b)$ be the basis of closure system obtained after replacing the original column of attribute b by its complement column $\neg b$. Then the *total support* of $\neg b$ can be computed, for each $a \in A \setminus b$, as before:

$$\operatorname{tsup}_{\neg b}(a) = \Sigma\{\frac{|sup(X)|}{|X|} : a \in X, (X \to \neg b) \in \beta(\neg b)\}.$$

Define now the parameter of relevance of parameter $a \in A \setminus b$ to parameter b, with respect to basis β:

$$\operatorname{rel}_b(a) = \frac{\operatorname{tsup}_b(a)}{\operatorname{tsup}_{\neg b}(a) + 1}.$$

The highest relevance of a is achieved by a combination of high total support of a in implications $X \to b$ and low total support in implications $X \to \neg b$. This parameter provides the ranking of all parameters $a \in A \setminus b$, but also allows us to rank implications $X \to b$ in the basis, by computing the average of $\operatorname{rel}_b(x)$ for $x \in X$:

$$\operatorname{rel}_b(X \to b) = \frac{\Sigma\{\operatorname{rel}_b(x) : x \in X\}}{|X|}.$$

We emphasize that while the measurement of relevance of an attribute a with respect to b can be done for any basis, there should be some assumption about irreducibility of the antecedents in implications. Indeed, for each implication $X \to b$ one may add to the basis another implication $X \cup \{s\} \to b$, for some fixed attribute s. In this new basis, the attribute s may obtain an unnecessarily high measurement. As we pointed earlier in Sec. 2, both the D-basis and the canonical direct unit basis have the property of irreducibility for antecedents in their implications.

It would be interesting to compare the measurement of relevance parameters for individual attributes with respect to various bases and check whether the group of attributes with the high ranking will be independent of the choice of the basis.

4 Illustrating example from test data on ovarian cancer

We will be illustrating our approach on a relatively small data set composed of genes found to be highly correlated with microRNA and DNA methylation in a common set of 291 serous ovarian tumor samples. Global gene expression, microRNA and DNA methylation data for each tumor sample were downloaded from The Cancer Genome Atlas (TCGA) along with meta-data that included censored time-to-death from all causes (survival) [18].

The resulting data matrices for each data type were jointly analyzed using matrix factorizations of rank-1 to identify a low-dimensional signature composed of genes, microRNA and DNA methylation loci that best represented the dominant source of variation in the data as a sparse linear model.

Hierarchical clustering and pathway analysis methods were then employed to identify an even smaller set of genes that continued to model the dominant signal as a sparse linear combination. We hypothesized that gene signatures obtained in this way would help to unravel the complex, inter-connected biology that drives the clinical trajectory of ovarian cancer. In particular, we focused on a gene expression signature composed of 21 genes (out of 16,000 interrogated) that were all direct down-stream targets of the OSM gene as determined by pathway analysis methods.

The gene expression profiles of the 21 genes were arranged in a binary table with 190 rows and 46 columns. The rows represent the ovarian cancer patients who participated in the study which observed their survival time for 2500 days after treatment with standard chemotherapy with cisplatin and paclitaxel.

The first 42 columns represent the indicator functions for the expression levels of the 21 genes (after quantile normalization). If patient y has relatively high expression of gene x, it will be marked by indicator 1 in column x, and when this patient shows relatively low levels of expression of gene x, the indicator value of 1 is put in column $21 + x$. Those patients whose gene expression is within some threshold around the average expression value in the group will have an indicator value of 0 in both columns x and $21 + x$.

The last four columns represent the survival groups within those 190 patients. Indicator is 1 in column 43 if a patient lived longer than 2000 days, and it is 1 in column 44 if she lived longer than 1300. Thus, the implication $43 \to 44$ holds in the binary table. The indicator is 1 in the 45th column if a patient lived less than 1300 days, and it is 1 in column 46 if she lived less than 850 days. Hence another implication $46 \to 45$ is also a part of the basis.

The cut-off thresholds for survival of 2000, 1300 and 850 days roughly correspond to quartiles for the entire group of 291 patients based on Kaplan-Meier analysis. Recall that the whole observation group was comprised of 291 patients, of which only 153 stayed in the study for the total period of 2500 days, while the remaining patients were observed for shorter periods. Nevertheless, 38 of them were observed long enough to include them into two upper quartiles, with some partial loss of information for those between 1300 and 2000 days of observation. A total of 101 patients were excluded from the testing related to survival for this test of D-basis extraction. These were patients who either left the study before 2000 days, or else had survived but for fewer than 2000 days at the end of the study, and hence could not be assigned to a survival cohort. Other ways of dealing with censoring that would include all the samples are discussed in the next section.

Let us illustrate the measurement of the implications in the D-basis of this matrix, when the target attribute b is 44, i.e., the indicator of longer surviving patients (at least 1300 days). In this particular study this is the group of 87 patients, which includes a subgroup of 31 patients who survived longer than 2000 days. Application of the algorithm from [1] with the request of $\beta(44)$ produces 1819 implications of the form $X \to 44$, including the expected implication $43 \to 44$.

It is possible to rank the implications in the retrieved basis by support. For example, among 1819 implications obtained, there is a single one with the highest cardinality of the support = 9: $\{16, 28\} \to 44$. (Here 16 represents high expression of the gene GBP2, while 28 is low expression of IL7.) There is also one with the support of 8: $\{14, 1, 3, 11\} \to 44$ where 14=HLA-B, 1=VDR, 3=TRIM22, 11=IL15. There are also 9 implications of support 7, and 17 implications of support 6. The great majority of implications have a support of 1 or 2, and it is hard to decide whether any implications from this large group could be of particular value.

With the new approach we were able to compute the relevance parameter for all the columns and choose the columns of the highest relevance. In our case, these were column 29 with relevance parameter 2.7894, column 9 with relevance value of 2.5137, and columns 1 and 4 with the relevance figures 1.8702 and 1.8425, respectively. (Column 29 is low expression of IL4R, while 9 is high expression of IL1B, 1=VDR, 4=SELE.)

The value 2.7894 for column 29 can be interpreted as following: attribute 29 appeared in implications with the conclusion $b = 44$, i.e., indicator that the patient was in the longer surviving group (patients with the survival period longer than 1300 days), approximately 2.7 times more often than for the complement of this group (patients surviving less than 1300 days). According to the definition of tsup_{44} parameter, the contribution of each individual implication would be adjusted by the weight $\frac{\sup(X)}{|X|}$. In the $\beta(44)$ section of the D-basis the size of antecedent varied between 2 and 7, with most of implications having 3 or 4 attributes in their antecedent.

In any outcome of the testing, the follow-up validation of the discovery assumes the check on an additional data set of 99 ovarian cancer patients. The identified groups of parameters highly relevant for survival are validated, when they successfully separate the survival curves for both training and test data, based on Kaplan-Meier and Cox regression analysis.

The six genes (from this set of 21 targets of OSM) with the highest relevance to long survival (over 1300 days) turned out to be IL4R, IL1B, VDR, SELE, HLA-B, GBP2 and IL15RA. We did a Kaplan-Meier analysis of the signature on the 291 patient sample, and then tested it on the independent sample of 99 other ovarian cancer patients. The difference between the KM plots for the top and bottom quartiles of the 291 training samples ordered by the 6-gene D-basis signature are statistically significant in both the KM analysis (p=0.00217) and Cox regression analysis (p=0.0000269). The same analysis on the 99 validation samples gave p=0.0176 for the KM analysis and p=0.0419 for the Cox regression analysis. Thus the six-gene signature derived from the relevance parameter is associated with survival at a significant level.

5 The larger test case of ovarian cancer

More comprehensive testing was done on a set of 40 genes that are downstream targets of IL4, identified by the combination of methods described in Sec. 4. This time all 291 patients were included into the testing, thus the size of the matrix was 291 × 84, where the first 80 columns represented the indicators for the high and low levels of expression for 40 genes, and the 4 columns represented 4 groups of patients based on the observed time to death due to all causes.

The 191 patients described in Sec. 4 had indicators placing them into longest, long, short and shortest surviving groups. The remaining 101 patients were coded by all zeros in 4 surviving groups columns. These patients were observed in the study for less than 1300 days, while their status after the last observation remained unknown: this is a group of so-called *censored* patients. Potentially, given longer term of observation, each patient from the censored group could appear in any of the surviving groups.

From the point of extraction of implications, inclusion of censored patients without marking them into survival groups results in the loss of a subset of implications that may belong to the D-basis. For example, if implication $X \to b$ holds in all the rows of 191 patients, with $b = 81$, which is a column of the longest surviving patients, and one of the patients from the group of 101, which was marked by all zeros, has all the parameters from X present, then the implication $X \to b$ fails in the row of matrix representing this patient, so that this implication is excluded from the output. Thus, the experiment on 291×84 matrix produces a subset of the basis for 190×84.

Another way to include the data of all 291 patients into the analysis is to assign 101 censored patients into 4 groups, based on the Kaplan-Meier analysis of 291 samples; see [13]. Potentially, marking censored patients into survival groups based on risk analysis may result in both excluding and adding some implications to the D-basis. This method will be used in some of our future tests.

We are planning to report on the full extent of this testing in a forthcoming publication. For the purposes of the current report, we outline the outcomes of the testing done on 291×84 matrix, when censored patients were marked by all zeros in the columns $81 - 84$, representing the survival groups.

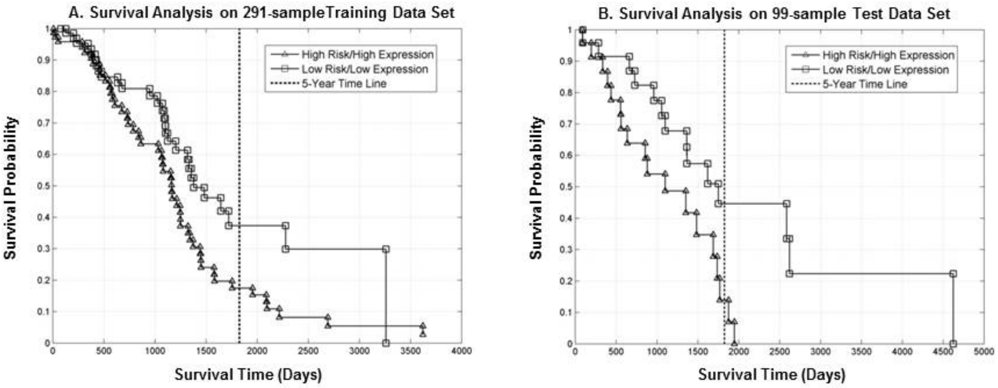


Figure 1. Kaplan-Meier (KM) analysis of training and independent test data stratified by a 6-gene *D*-base expression signature (*DBSig6*). **Panel A**. KM plots of top and bottom quartiles of 291 ovarian tumors ordered by the 6-gene *DBSig6* expression signature. The KM plot marked by squares models the survival of the lowest quartile of patients ordered by *DBSig6* expression and KM plot of up-triangles models patients in highest quartile. The difference in KM plots is statistically significant (logrankP = 0.0143) and *DBSig6* expression is still associated with survival after adjustment for age and stage (CoxP = 0.0011). The intersection of the vertical dashed line with each KM curve gives the 5-year survival rate for each group of patients on the vertical axis. **Panel B**. KM plots on independent test data set composed of 99 ovarian tumros. The interpretation of the KM plots marked by squares and up-triangles and the vertical dashed line are the same as in Panel A. The difference in KM plots is statistically significant (logrankP = 0.0137) and *DBSig6* remains predictive of survival even after adjustment for age and stage (CoxP = 0.0086). Panels A and B demonstrate that the *DBSig6* expression signature is able to robustly identify "good" and "bad" responders to standard chemotherapy for ovarian cancer.

The D-basis was extracted with 4 requests, for $b = 81$, 82, 83 and 84. We used a feature of the algorithmic design, which may request only a portion of the basis, selecting the implication with minimum support parameter "minsup." When minsup $=$ k, the program outputs only implications with the minimum support at least k. This considerably shortens the running time and the list of implications. For example, with $b = 83$ the algorithm produced 4325 implications of minimum support 3, in 91.94 sec.

The six genes (from this set of 40 targets of IL4) with the highest relevance to long survival (over 1300 days) turned out to be FCGR2A, CD86, IFI30, CCL5, SELPLG and ICOS, all of which are associated with immune response and cancer. The results of the Kaplan-Meier and Cox regression analysis of this six-gene signature associated with IL4, on both the 291 patient training set and 99 sample validation data, are shown in Fig. 1. The difference between the KM plots for the top and bottom quartiles on the training data ordered by the 6-gene D-basis signature are statistically significant in both the KM analysis (p=0.0143) and Cox regression analysis (p=0.00112). The same analysis on the 99 validation samples gave p=0.0137 for the KM analysis and p=0.00858 for the Cox regression analysis. Again, this six-gene signature derived from the relevance parameter is associated with survival at a significant level.

On the other hand, the genes relevant to short survival did not separate the curves significantly, nor did the combined sets of genes for long and short survival. This may be an artifact of how we dealt with censoring. Other tests indicate that the 40-gene set associated with IL4 is much richer than the 21-gene set derived from OSM, and we anticipate that this will show up in further relevance experiments.

6 Analysis of temporal data for the patients with brain tumor

The original set of data for 61 patients with brain tumors (astrocytomas, glioblastomas, and meningiomas) under new regimen of treatment was collected over the two years of observation in the hospital of Medical Holding in Astana, between 2012–2013. Patients were accepted into the experimental group after all options of standard chemotherapy were exhausted.

Three groups of parameters were regularly measured in patients. The first group was set of flow cytometry markers to identify major immune cell populations in peripheral blood: T helper cells, cytotoxic T cells, natural killers, B cells, and antigen-presenting cells. The second group was blood analysis for creatinine, bilirubin, calcium, protein, amylase, transferrin, C-reactive protein, immunoglobulins, lipase and iron. Finally, the third group of parameters was infectious markers including indicators for hepatitis A and C, cytomegalovirus, chlamydia, herpes simplex, Epstein-Barr virus, mycoplasma, ureaplasma, echinococcus, Helicobacter pylori, toxoplasma, and rubella.

The main read-out measurement of patient response to the new treatment was clinical assessment accompanied by the immune parameters, blood biochemistry parameters and the dynamic of infections.

All patients were divided into four groups depending on a clinical assessment. The first group included patients who did not survive cancer. The second group included patients succumbing to the illness. The third group included patients with stable health/tumor status. Finally, the fourth group incorporated patients with improving clinical assessments accompanied by reducing volume of tumor.

The goal of the study was to identify the elements of the treatment and measured parameters, which are associated to the positive response to the treatment.

The challenge of this data set was in the fact that each parameter was measured multiple times during the course of the treatment, while treatment and patient survival were of different time spans, resulting in significant variation in number of measurements. In order to take into consideration the dynamics of multiple parameters, and be able to analyze kinetics of very different length in similar consistent way, we transformed raw data into a set of increments, or differentials.

If the initial value of a parameter is V1, the value at the middle of the observation period is V2, and the final value during observation is V3, then the increment δ_{31} is defined as (V3-V1)/V1; increment δ_{21} is (V2-V1)/V1; increment δ_{32} is (V3-V2)/V2; Avg (average) is calculated average value for the parameter during the entire observation period. The analysis is performed in terms of (V1, Avg, $\delta_{21}, \delta_{32}, \delta_{31}$).

For conversion of real-valued data into binary form, the initial values V1 were taken into consideration only for part of the parameters. The full range of V1 parameter values was divided into quartiles and were coded into 4 columns representing the quartiles. The dynamic parameters were converted into two columns each, where the mark 1 in one of columns was an indicator of increasing of this parameter more than 10% from average variation within the group, over the indicated period of treatment, while 1 in the other column was the indicator of decreasing by more than 10%. Both columns would be marked by 0 when the parameter stayed within 10% of average variation.

The 4 clinical assessment groups were combined into two: C1 group comprised the two groups of declining patients, and C2 group the stabilizing or improving patients.

The resulting binary table included 287 columns for 61 patients. An additional table was created for a subgroup of 33 patients with identical diagnosis of specific brain cancer, while the whole group incorporated patients with different sub-types of brain tumors.

The request for computation of the basis for a column that combined stabilizing and improved patients (C2) resulted in 1,138,518 implications computed in 39639 sec, or just over 11 hours. For the column indicating the group of declining patients (C1) the number of implications was 2,073,282, and it was computed in 170458 sec, or 47.34 hours.

The computation of the relevance parameter for all attributes, with respect to columns C1 and C2, provided the ranking of attributes in each case. For the computation of the most relevant implications, it was considered reasonable to make another run of the program to filter the implications first with respect to the minimum support parameter. It was established at the level minsup = 5, so that only those implications were produced whose antecedent held for at least 5 patients. The test for C1 now took only 1400 sec and produced only 9,794 implications. The test for C2 took 345 sec and produced 19,112 implications.

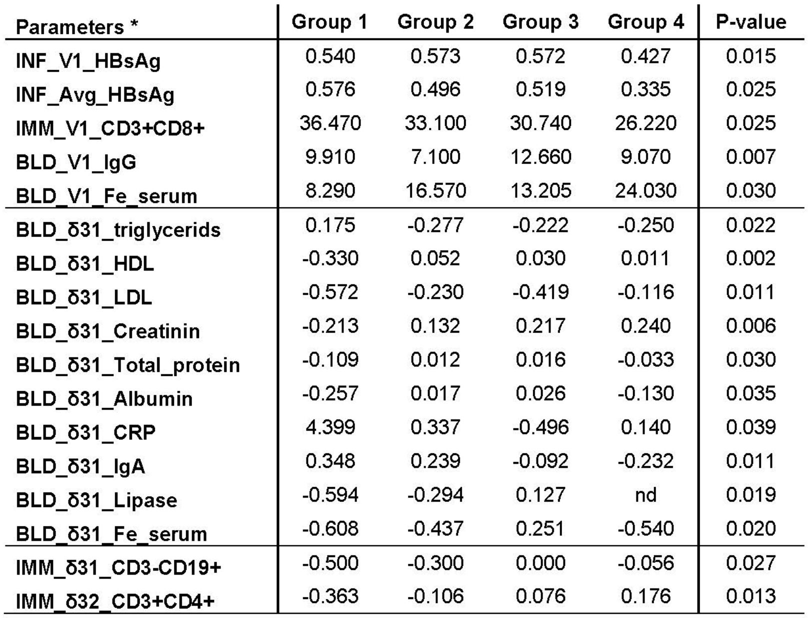
Similar tests were reproduced for the sub-group of 33 patients with the specific diagnosis of brain cancer. We observed much higher variation of the relevance parameter in the case of 61 patients. For example, several attributes showed relevance in the range of 1,000-10,000, mostly in the cases when $\mathrm{tsup}_{\neg b}(a)$ for the attribute a had 0 value. Most of them got the relevance below 1 in the test of the sub-group of 33.

At the same time, most of the highly relevant attributes in the test on the subgroup of 33 patients showed their significance in the test for 61 as well.

On the set of 33 patients, the ranking of attributes by the relevance to column C1 revealed the attributes in ranking positions 1, 2 and 7, which correspond to dynamics of the same immune parameter: CD3+CD8+ cytotoxic T cells in the first, second halves of the treatment, and during the whole period, respectively. The dynamics was decreasing at the start, increasing in the second half, but still decreasing overall. In the C2 group, highly relevant were attributes in ranking positions 7 and 9 were decreasing dynamics of the presence of two specific viruses.

Statistical analysis of the immune, biochemical and infection parameters: initial values, averages and increments (V1, Avg, δ_{21}, δ_{32}, δ_{31}) was also performed using non-

Table 2. Parameters associated with patients' clinical assessment.



Parameters *	Group 1	Group 2	Group 3	Group 4	P-value
INF_V1_HBsAg	0.540	0.573	0.572	0.427	0.015
INF_Avg_HBsAg	0.576	0.496	0.519	0.335	0.025
IMM_V1_CD3+CD8+	36.470	33.100	30.740	26.220	0.025
BLD_V1_IgG	9.910	7.100	12.660	9.070	0.007
BLD_V1_Fe_serum	8.290	16.570	13.205	24.030	0.030
BLD_δ31_triglycerids	0.175	-0.277	-0.222	-0.250	0.022
BLD_δ31_HDL	-0.330	0.052	0.030	0.011	0.002
BLD_δ31_LDL	-0.572	-0.230	-0.419	-0.116	0.011
BLD_δ31_Creatinin	-0.213	0.132	0.217	0.240	0.006
BLD_δ31_Total_protein	-0.109	0.012	0.016	-0.033	0.030
BLD_δ31_Albumin	-0.257	0.017	0.026	-0.130	0.035
BLD_δ31_CRP	4.399	0.337	-0.496	0.140	0.039
BLD_δ31_IgA	0.348	0.239	-0.092	-0.232	0.011
BLD_δ31_Lipase	-0.594	-0.294	0.127	nd	0.019
BLD_δ31_Fe_serum	-0.608	-0.437	0.251	-0.540	0.020
IMM_δ31_CD3-CD19+	-0.500	-0.300	0.000	-0.056	0.027
IMM_δ32_CD3+CD4+	-0.363	-0.106	0.076	0.176	0.013

* **BLD** – patients' blood parameters, **IMM** – immune parameters, **INF** – infection parameters. **HBsAg** - hepatitis B virus surface antigen, **CD3+CD8+** - cytotoxic T cells levels in peripheral blood, **IgG** – total immunoglobulins, **Fe_serum** - iron in blood, **HDL** – high density lipoproteins, **LDL** – low density lipoproteins, **CRP** - C-reactive protein, **IgA** – immunoglobulin of IgA isotype, **CD3-CD19+** - level of B cells in peripheral blood, **CD3+CD4+** - level of T helper cells. **Group** –median value of the parameter is presented for each group of patients. **P-value** – statistical significance of the Kruskal-Wallis non-parametric test.

parametric Spearman's correlation analysis [17] to find ties between parameters, see analogous use of increments in Adarichev et al. [3, 4]. Global cross-correlation of all real-valued increments was performed for the data on 61 patients. Using analysis of correlation, the biases between only two parameters at a time could be studied, which is a major nuisance over the implication approach. For analysis of the parameters' difference between groups of patients, we used the non-parametric Kruskal-Wallis test in the R language for statistical computing [19]. This test is an equivalent of the one-way analysis of variance (ANOVA), but does not require assumption of the normal distribution of data.

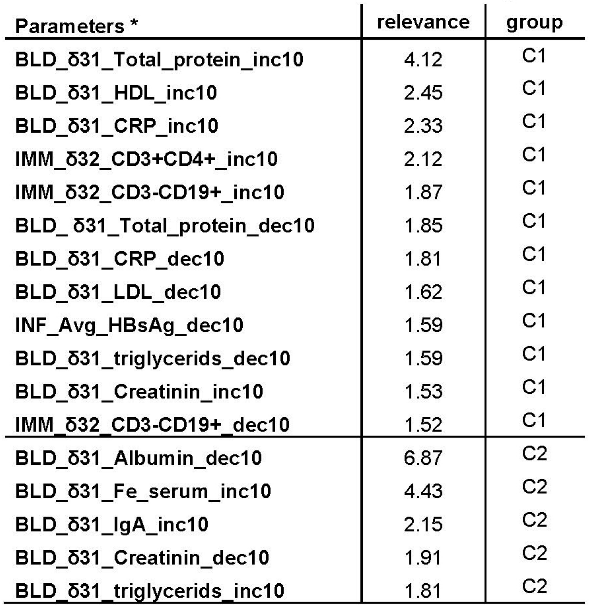
Initial parameters (V1) most significantly associated with clinical assessment were HBsAg surface antigen of the hepatitis B virus ($p < 0.015$), amount of CD3+CD8+ cytotoxic T cells in peripheral blood ($p < 0.025$), total IgG immunoglobulins ($p < 0.007$), and higher concentration of blood iron ($p < 0.03$), see Table 2.

Change of blood parameters over the entire period of observation (δ_{31}) produced the longest list of significant associations: lipoproteins (triglycerides, HDL, LDL), creatinine, total protein, albumin, C-reactive protein, immunoglobulin IgA, lipase, and blood iron, refer to Table 2.

Importance of the total IgG immunoglobulins and specifically IgA isotype was in line with levels of B cells in peripheral blood (Table 3, $p < 0.02$), cells that produce immunoglobulins. Increase of T helper cells levels significantly correlated with good prognosis for survival ($p < 0.01$), see Table 2.

There were 151 attributes with the relevance above 1.5 associated with attribute C1, and 25 attributes associated with C2 with the same relevance threshold. Parallel analysis of same dataset using Kruskal-Wallis statistical test discovered 27 parameters at $p < 0.05$ significance level, they are presented in Table 3.

Table 3. Parameters discovered with both implications and statistical approaches.



Parameters *	relevance	group
BLD_δ31_Total_protein_inc10	4.12	C1
BLD_δ31_HDL_inc10	2.45	C1
BLD_δ31_CRP_inc10	2.33	C1
IMM_δ32_CD3+CD4+_inc10	2.12	C1
IMM_δ32_CD3-CD19+_inc10	1.87	C1
BLD_ δ31_Total_protein_dec10	1.85	C1
BLD_δ31_CRP_dec10	1.81	C1
BLD_δ31_LDL_dec10	1.62	C1
INF_Avg_HBsAg_dec10	1.59	C1
BLD_δ31_triglycerids_dec10	1.59	C1
BLD_δ31_Creatinin_inc10	1.53	C1
IMM_δ32_CD3-CD19+_dec10	1.52	C1
BLD_δ31_Albumin_dec10	6.87	C2
BLD_δ31_Fe_serum_inc10	4.43	C2
BLD_δ31_IgA_inc10	2.15	C2
BLD_δ31_Creatinin_dec10	1.91	C2
BLD_δ31_triglycerids_inc10	1.81	C2

Overlapping of these two analyses was compared using a Venn diagram. Twelve parameters were found both by implication and statistical approaches in C1, and five parameters were discovered in C2. Details of parameters are presented in Fig. 2.

The results presented in Tables 2 and 3 could be further corroborated by biological functionality of the discovered parameters. Parameter IMM_δ32_CD3-CD19+ reflects number of CD3-CD19+ B cells that produce antibodies. Correspondingly, blood parameter BLD_δ31_IgA is also in the list in the C2 group. Armed T cells, which are represented with IMM_δ32_CD3+CD4+ parameter could actually stimulate B cells to produce antibodies. Another group of parameters is related to blood lipoproteins of high and low density (HDL and LDL, respectively) and triglycerids. These parameters are known to be biased in the norm and pathology. Brain tumorigenesis and pharmaceutical intervention to fight cancer both lead to inflammatory reactions. Correspondingly, we found inflammatory marker C-reactive protein in patients blood using both methods (parameter BLD_δ31_CRP). In brief, this set of parameter is functionally valid for the pathology under investigation. In the test on the binary conversion of this data, all blood parameters from the table showed high relevance of increasing trend for group C1, and four of them high relevance of decreasing trend in group C2, which confirms statistical observation.

7 Concluding remarks and future research

This paper comes on the heels of a series of discoveries about computational complexity of extraction of the canonical basis of Guigues-Duquenne; see [6] and [10]. With the D-basis and algorithm for extracting implications derived in [1], the problem of handling

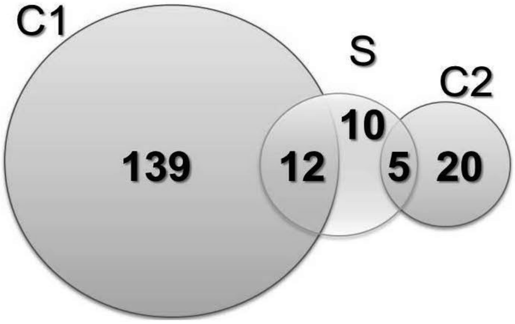


Figure 2. Venn diagram for overlapping results of implication and statistical approaches. Circle **C1** - group of declining patients with total of 139+12=151 implications. Circle **C2** – group of improving patients with total of 20+5=25 implications. Threshold for relevance is set at 1.5. Circle **S** – results of statistical analysis with total of 12+10+5= 27 significant biases at $p < 0.05$ threshold. See Tables 2 and 3 for details.

relatively large data sets, which may include hundreds of attributes and objects, can be considered tamed. This brings researchers to a new challenge, already battled in data mining: the output of algorithms producing staggering amounts of association rules, for which further methods of analysis and filtering are needed.

Association rules were brought to consideration in analysis of transaction data. While the choice of purchases follow particular patterns, these are rather "soft" rules which do not need to hold even in a majority of all transactions. Whenever we come to analysis of biological data, the dependencies between attributes may reveal the laws of nature which are yet to be discovered. Thus, one may hope to find higher confidence levels of association rules discovered in this type of data.

The association rules of highest confidence (=1) are implications, and extraction of this special subset of association rules allows different approaches based on underlying structure of closure operators. On the other hand, the main metric for association rules, the confidence, is no longer a player for selection of important implications. This requires the development of new measurements for implications that would allow us to restrict our attention to those that may discover the hidden laws.

The main achievement of the current paper is the introduction of a new measurement for implications, which we call the *relevance*. It enables us to rank the attributes with respect to some fixed attribute b, in some given basis of implications, then apply individual relevance values to compute the relevance of implications. In particular, we can use this to measure the relevance of genetic or medical data to clinical outcomes.

The new approach was tested on two sources of medical data related to clinical assessment or survival of cancer patients. Our initial testing has already shown that the relevance parameter can be used to find genetic signatures associated with longer survival of ovarian cancer patients. The full analysis of these data sources will continue and we plan tuning the computation of relevance metric after validation of results of testing on additional data sets.

Acknowledgements. The C++ code for D-basis extraction on the binary table input used for testing in this project was created by undergraduate students of Yeshiva College in New York: Joshua Blumenkopf and Toviah Moldvin. We received the permission of Takeaki Uno, from the National Institute of Informatics in Tokyo, to implement the call to his subroutine performing the hypergraph dualization, within the structure of our programming code. We were assisted by colleagues Ulrich Norbisrath and Mark Sterling, from the Computer Science Department of School of Science and Technology of NU, when we needed tuning and debugging of the code, also to Rustam Bekishev and Anel Nurtay for assistance in the project. The first author is grateful to the bio-informatics group of the University of Hawaii Cancer Center, for the welcoming atmosphere and fruitful collaboration during her visit in June 2014, supported by Nazarbayev University grant N 13/42. The second author expresses his gratitude for support of his visit to Nazarbayev University in May–June 2013 and May 2014, which were partly funded by NU grant N 13/42 and grant N 0112PK02175 of Medical Holding of Astana. Tom Wenska, Ashkan Zeinalzadeh and Jenna Maligro contributed to the research and discussion in Honolulu.

References

1. K. Adaricheva, and J.B. Nation, *Discovery of the D-basis in binary tables based on hypergraph dualization*, submitted to Theoretical Computer Science.
2. K. Adaricheva, J.B. Nation and R. Rand, *Ordered direct implicational basis of a finite closure system*, Disc. Appl. Math. **161** (2013), 707–723.
3. V.A. Adarichev, C. Vermes, A. Hanyecz, K. Mikecz, E.G. Bremer, and T.T. Glant, *Gene expression profiling in murine autoimmune arthritis during the initiation and progression of joint inflammation*, Arthritis Research and Therapy **7** (2005), 196–207.
4. V.A. Adarichev, C. Vermes, A. Hanyecz, K. Ludanyi, M. Tunyogi-Csapó, K. Mikecz, and T.T. Glant, *Antigen-induced differential gene expression in lymphocytes and gene expression profile in synovium prior to the onset of arthritis*, Autoimmunity **39** (2006), 663–673.
5. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, *Fast discovery of association rules*, Advances in Knowledge discovery and data mining, AAAI Press, Menlo Park, California (1996), 307–328.
6. M.A. Babin and S.O. Kuznetsov, *Computing premises of a minimal cover of functional dependencies is intractable*, Disc. Appl. Math. **161** (2013), 742–749.
7. J.L. Balcázar, *Redundancy, deduction schemes, and minimum-size bases for association rules*, Log. Meth. Comput. Sci. **6** (2010), 2:3, 1–33.
8. K. Bertet and B. Monjardet, *The multiple facets of the canonical direct unit implicational basis*, Theoretical Computer Science **411** (2010), 2155–2166.
9. E. Boros, K. Elbassioni, V. Gurvich and L. Khachiyan, *Generating dual-bounded hypergraphs*, Optimization Methods and Software, **17** (2002), 749–781.
10. F. Distel and B. Sertkaya, *On the complexity of enumerating the pseudo-intents*, Disc. Appl. Math. **159** (2011), 450–466.
11. M. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive normal forms*, J. Algorithms **21** (1996), 618–628.
12. J. L. Guigues and V. Duquenne, *Familles minimales d'implications informatives résultant d'une tables de données binares*, Math. Sci. Hum. **95** (1986), 5–18.
13. E.L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, J. Amer. Statist. Assn. **53** N282 (1958), 457–481.
14. M. Kryszkiewicz, *Concise representation of association rules*, Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery, Springer-Verlag, London, UK, 92–109.
15. K. Murakami and T. Uno, *Efficient algorithms for dualizing large scale hypergraphs*, Disc. Appl. Math. **170** (2014), 83–94.
16. U. Ryssel, F. Distel and D. Borchmann, *Fast algorithms for implication bases and attribute exploration using proper premises*, Ann. Math. Art. Intell. **70** (2014), 25–53.
17. C. Spearman, *The proof and measurement of association between two things*, Amer. J. Psychol. **15** (1904), 72–101.
18. The Cancer Genome Atlas Research Network, *The Cancer Genome Atlas Pan-Cancer analysis project*, Nature Genetics **45** (2013), 1113–1120.
19. R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.