

Combinatorial Analysis of Breast Cancer Data from Gene Expression Microarrays

Gabriela Alexe¹, Sorin Alexe¹, David Axelrod², Endre Boros¹, Michael Reiss³, Peter L. Hammer¹

Abstract Using the methodology of the *Logical Analysis of Data (LAD)* we have re-analyzed the publicly available breast cancer gene expression microarray dataset (<http://www.rii.com/publications/2002/vantveer.htm>) studied by van't Veer et al. in 2000. The two main motivations for reexamining this dataset using *LAD* were (i) to evaluate the accuracy of a *LAD*-based prognostic system, and (ii) to derive additional conclusions about the problem. After a brief introduction to *LAD*, we present a set of 16 genes, not all of which are highly correlated with the outcome (metastasis within 5 years), whose collective power of distinguishing positive and negative cases allows the construction of a highly accurate *LAD*-based prognostic system. Applying the evaluation measure used in the van't Veer dataset, the accuracy of this *LAD* classifier on the training and the test sets turns out to be of 100% and 94.7%, respectively. The construction is fully reproducible, and the proposed prognostic system provides, along with the classification of any new case, a clear explanation of the reasons for its classification. The patterns (collective biomarkers) identified by *LAD* are shown to provide additional conclusions about patients (e.g., identification of two new classes of patients with highly distinguished features) and genes (including the identification of several contributor or inhibitor genes). Finally, a *LAD*-based analysis suggests the existence of dissimilarities between the cases in the training and those in the test set.

1. INTRODUCTION

Using the methodology of the *Logical Analysis of Data (LAD)* ([11], [12]) we have re-analyzed the publicly available microarray dataset (<http://www.rii.com/publications/2002/vantveer.htm>) studied by van't Veer et al. in [17]. The main goal of van't Veer's study was to predict the clinical outcome of breast cancer (i.e., to identify those cases which will develop metastases within 5 years) based on the analysis of gene expression signatures. The crucial importance of this problem is due to the fact that the available adjuvant (chemo or hormone) therapy, which reduces by about one third the risk of distant metastases, is not really necessary for 70-80% of the patients who currently receive it; moreover, this therapy can have serious side effects, and involves high medical costs. The study [17] illustrates clearly that machine learning techniques, data mining, and other new techniques applied to DNA microarray analysis can outperform most clinical predictors currently in use for breast cancer. The study concludes that the new findings “provide a strategy to select patients who would benefit from adjuvant therapy”.

A specific feature of datasets coming from genomics is the presence of a very large number of measurements concerning gene expressions, but only a relatively small number of observations. For instance, the attributes in the van't Veer study correspond to more than 25,000 human genes, while the number of cases is only 97. In this dataset, each case is described by the expression levels of 25,000 genes as measured by fluorescence intensities of RNA hybridized to microarrays of oligonucleotides. The cases in the dataset represent 97 lymph-node-negative breast cancer patients, who are grouped into a *training set* of 78, and a *test set* of 19 cases. The training set includes 34 *positive* cases (having a “poor prognosis” signature, i.e., having less than 5 years of metastasis-free survival), and 44 *negative* cases (having a “good prognosis” signature, i.e., having more than 5 years of metastasis-free survival). The test set includes 12 positive and 7 negative cases.

¹ RUTCOR - Rutgers University's Center for Operations Research, Piscataway, NJ 08854, USA, {alex, salex, boros, hammer}@rutcor.rutgers.edu

² Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA, Axelrod@nel-exchange.Rutgers.Edu

³ Division of Medical Oncology, UMDNJ-Robert Wood Johnson Medical School, The Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA, michael.reiss@UMDNJ.EDU

The van't Veer study used DNA microarray analysis on primary breast tumors, and “applied supervised classification to identify gene expression signature strongly predictive of a short interval to distant metastases (‘poor prognosis’ signature) in patients without tumor cells in local lymph nodes at diagnosis (lymph node negative)”. The study identified as significant markers of metastases 231 genes, all of whose correlations with the outcome exceeded 0.3 in absolute value, and constructed an optimal prognosis classifier based on the best 70 ones. The classifier predicted with an 83.33% accuracy (65 correct predictions out of 78) the positive or negative nature of the cases in the training set, and with an 89.47% accuracy (17 correct predictions out of 19) the classification of the test cases.

The present study is based on *LAD*, a combinatorics, optimization, and logic-based methodology for the analysis of data. Specific features of the *LAD* approach include the exhaustive examination of the entire set of features (without the exclusion of those having low statistical correlations with the outcome, or those having low expression levels), focusing on the classification power of the combinations of features (without confining attention only to the individual ones), and on the possibility of extracting novel information on the role of features and of combinations of features through the analysis of these exhaustive lists.

LAD has been shown to offer important insights in problems ranging from oil exploration ([11]), labor productivity analysis ([14]), country creditworthiness evaluation ([9]), to medical application (e.g., risk evaluation among cardiac patients ([7], [16])), polymer design for artificial bones ([1]). Recently ([4]), a highly accurate ovarian cancer diagnostic system was developed on the basis of a proteomic dataset.

The present study is the first attempt of using *LAD* for the analysis of a genomic dataset. The two main reasons for re-examining the [17] dataset using *LAD* were (i) to evaluate the potential of *LAD* in developing a prognostic system for breast cancer using genomic data, and (ii) to derive additional conclusions about the nature of some of the genes, about special classes of patients, etc.

After a brief introduction to *LAD*, we present a set of 16 genes (not all of which are highly correlated with the outcome), whose collective power of distinguishing positive and negative cases allows the construction of a highly accurate *LAD*-based prognostic system. Applying the evaluation measure used in [17], the accuracy of this *LAD* classifier on the training and the test sets turns out to be of 100% and 94.7%, respectively. The construction is fully reproducible, and the proposed prognostic system provides along with the classification of any new case, a clear explanation of the reasons for its classification. The patterns (collective biomarkers) identified by *LAD* are shown to provide additional information on the influence and nature of the 16 genes, and to define two new classes of patients with highly distinguished features. Finally, a *LAD*-based analysis suggests the existence of dissimilarities between the cases in the training and those in the test set.

2. MATERIALS AND METHODS

2.1. Basic Concepts of LAD

It can be expected that “large” or “small” values of the expression levels of certain genes can determine the poor or bad prognosis of a breast cancer patient. In order to express such relations in more precise terms, it is natural to replace terms like “large” or “small” by conditions of the type “...is more than” or “...is less than” a certain value. It is therefore natural to examine the role of well chosen *cutpoints* associated to the expression levels of the genes. For instance, the imprecise observation that low intensity levels of gene *NM_014675* are (more or less) characteristic for a poor prognosis, can be reformulated as the ultra-simplistic classification system “*If the intensity level of the gene NM_014675 is less than -0.05 then the patient has a poor prognosis*”, the assumption of which is valid for 19 positive and 8 negative cases in the training set, i.e., it has a sensitivity of 70.4%, and a specificity of 70.6%).

Combinations of such cutpoint-based conditions extend naturally this idea. For instance, the combined requirement of satisfying the two conditions “*The intensity level of the gene NM_014675 is less than -0.05*” and “*The intensity level of the gene NM_005243 is less than -0.028*” is fulfilled by 11

of the 34 positive cases in the dataset, and by none of the negative ones. Again, these two requirements could be viewed as a classification system of poor prognosis cases, having a sensitivity of 100%, and a specificity of 65.7%.

Such ideas are at the foundation of the *Logical Analysis of Data (LAD)*. The essence of *LAD* is (i) to detect *patterns*, or *collective biomarkers*, i.e., simple classifiers consisting in restrictions imposed on the values of the expression levels of the intensities of a combination of several genes, (ii) to exhaustively *generate patterns* in an algorithmically efficient way, (iii) to use the collection of patterns as a *prognostic system* and thoroughly validate it, (iv) to extract from this collection as much additional information as possible about the *role and nature of genes* in the dataset (i.e., to detect *contributors* and *blockers*), (v) to study the common characteristics of *groups of patients* which satisfy similar patterns.

We shall describe below the basic concepts used in *LAD*, including some of its computational aspects. In particular, we shall describe more precisely the concepts of *support sets*, *patterns*, *pandecks*, and *LAD-based classification systems*, and will discuss the *validation* techniques used.

2.1.1. Cutpoints and Binarization One of the underlying principles of *LAD* is to disregard the exact values of a variable (e.g., a gene), specifying for each patient only the information that the corresponding value of this variable is “large” or “small”. The implementation of this principle requires the determination of 2 *cutpoints* $c'_j \leq c''_j$ for the intensity levels of each gene j , such that expression levels of the gene intensity falling below c'_j are considered “low”, and those above c''_j are considered high. Let us remark first that in some problems the two cutpoints corresponding to a variable may coincide. Let us also remark that in certain problems, 2 cutpoints are not sufficient for distinguishing the positive observations from the negative ones ([10]).

LAD associates to each variable x_j and each possible cutpoint c_j a *binary* variable y_j equal to 1 whenever $x_j \geq c_j$, and to 0 otherwise. In this way, a numerical variable (e.g., specifying the expression levels of the intensity of a gene j) is transformed into a large number of binary variables. Since the size of the dataset – which has been very large from the beginning – increases even further, this problem is handled by carrying out a “filtering” process, which retains only a “support set” consisting of a very small number of these variables.

2.1.2. Support Sets In order to distinguish the measurements of good and of poor prognosis patients, only a tiny fraction of the information contained in the (original or binarized) dataset is needed. In particular, all the information about the vast majority of the genes in the dataset is redundant. At the same time, even for the genes which are not redundant, at most 2 of the corresponding binary variables are needed. A set of binary variables which are sufficient to distinguish poor and good prognosis cases will be called a *support set*. A support set is called *minimal* if none of its proper subsets is a support set; clearly, not every minimal support set is of minimum size. It is important to notice that a dataset may admit hundreds or thousands of minimal support sets. The reduction of a large dataset to a substantially smaller one, which includes only the variables in the chosen support set allows a major simplification of the problem, and has a great importance for diagnosis (although in some cases, the presence of a *limited number* of redundant variables may be acceptable for assuring a higher stability of the results).

The problem of finding minimal support sets was modeled in ([10], [11], [12]) as a typical “set-covering” problem, for the solution of which there are numerous methods known in combinatorial optimization. In our case, the excessive dimensions of the associated set-covering problem (approximately 20,000 constraints involving between 2 and 3 million 0-1 variables) required the use of powerful heuristics to trim down the size of the problem. In order to be able to handle the large problems typical for genomic and proteomic datasets, a general heuristic size-reduction procedure was developed in [5]. The essence of this method is to balance the conflicting criteria of minimizing size and maximizing discrimination between positive and negative observations. In contrast to many statistically-

based methods, the support set generation procedures of *LAD* are guided by the collective strength of the subsets of variables (peptides), without being necessarily restricted to those peptides which have the highest individual correlation coefficients with the outcome.

By applying this method to the breast cancer dataset analyzed in this paper, we identified a support set consisting of 32 binary variables associated to only 16 of the 25,000 genes. The high sensitivity and specificity of the prognostic system built on these 16 genes are due to a large extent to the qualities of the underlying support set.

2.1.3. Logical patterns A *conjunction* is a set of conditions requiring that the binary variables appearing in a selected subset of the support set take specific (0 or 1) values, i.e., that the expression levels of the corresponding genes should be below or above certain cutpoints. The typical conjunctions appearing in most data analysis studies fix the values of not more than 2 or 3 binary variables. A conjunction is called a *positive (negative) pattern* if its set of conditions are satisfied simultaneously by “sufficiently many” of the positive (negative) cases, and by “sufficiently few” of the negative (positive) cases.

For example, in the van't Veer breast cancer dataset, if “sufficiently many” is defined as “at least 30%”, then the two conditions "*The intensity level of the gene NM_014675 is less than -0.05*" and "*The intensity level of the gene NM_005243 is less than -0.028*" are fulfilled by 11 of the 34 positive cases in the training set, and by none of the negative cases; therefore, the simultaneous fulfillment of these two conditions describes a positive pattern, to be denoted below *P1*. Similarly, the two conditions "*The intensity level of the NM_016448 is less than -0.098*" and "*The intensity level of the gene AF131819 is less than 0.0205*" are fulfilled by 17 of the 44 negative cases in the dataset, and by none of the positive cases; therefore, the simultaneous fulfillment of these two conditions describes a negative pattern, to be denoted below *N1*.

Two of the most important characteristics of a pattern are its degree, and its coverage. The *degree* of a pattern is simply the number of variables (genes) involved in its defining conditions. In our example, both *P1* and *N1* have degree 2. A case *C* is said to *display* a pattern, or to be *covered* by it, if the corresponding intensity levels of the gene expressions satisfy the defining conditions of that pattern. The *prevalence* of a positive, respectively negative, pattern is simply the proportion of positive, respectively negative cases covered by it. For example, the two defining conditions of *P1* are satisfied simultaneously by 11 of the 34 positive cases, i.e., the prevalence of *P1* is 32.3%. Similarly, *N1* covers 17 of the 44 control cases, i.e. its prevalence is 38.6%. Patterns which cover *only* positive, or *only* negative cases are called *pure* patterns. Clearly, both *P1* and *N1* are pure patterns. Usually, datasets admitting pure patterns of low degrees and high prevalences allow the construction of reliable *LAD* diagnostic and prognostic systems.

Several combinatorial algorithms ([6], [8], [15]) have been recently developed for the efficient generation of libraries of patterns. As an indication of their efficiency, we mention that the generation of the 133,920 potential patterns examined for this study and the selection of the 132 maximal pure patterns require the total computer time of 5.1 seconds.

2.1.4. Pandects The *pandect*, i.e., the collection of all the positive and negative patterns corresponding to a dataset, is the major tool used in the *Logical Analysis of Data*, since it allows the construction of diagnostic and prognostic systems, the analysis of the importance and role of variables, the identification of new classes of observations, etc. In view of the enormous number of patterns corresponding to a dataset, the construction of the entire pandect is not realistic. However, it has been seen in numerous case studies that the knowledge of special subsets of the pandect is amply sufficient for the accurate analysis of datasets. The set of all positive (negative) patterns of degree at most d^+ (respectively, d^-) and prevalence at least p^+ (respectively, p^-) is called the (d^+, p^+) - *positive pandect* (respectively, the (d^-, p^-) - *negative pandect*). The best pandect-defining parameters d^+ , d^- , p^+ , p^- for the analysis of a

particular dataset are determined experimentally by carrying out a series of k -fold cross-validation experiments. The particular pandect used in the present study is defined by $d^+ = d^- = 3$, $p^+ = p^- = 35\%$, and consists of 39 positive and 93 negative patterns. While patterns can be viewed as *tests* indicative of a good or bad prognosis, the “pandect” plays the role of a high powered prognostic *battery of tests*. Clearly, the pandect is not a minimal system, since it may contain many redundant patterns, without which the system can still remain accurate. However, the built-in redundancy of the system can substantially increase ([3]) its “stability” or “robustness”, when applied to new cases.

2.1.5. Pattern-Space

In the given dataset, each patient is described in terms of approximately 25,000 attributes (genes) by specifying their respective expression levels. Taking into account that the *LAD* patterns can be viewed as logically synthesized attributes which can be expected to reflect more closely the condition of a patient than the original “raw data”, it is reasonable to assume that a description of patients in terms of the set of patterns she satisfies, should represent more precisely the patient’s condition. This *pattern-based representation* of the observations can be achieved by associating to each patient and to each pattern in the pandect an indicator variable showing whether the patient satisfies (indicator = 1) or does not satisfy (indicator = 0) the conditions describing that pattern. In this way, each patient is characterized by a sequence of zero-one values of the indicator variables associated to the positive and negative patterns in the pandect.

2.2. Prognostic *LAD* Systems: Construction and Validation

Several *LAD* prognostic systems were constructed for the dataset in [17]; we report below the steps in the construction of the final system. The parameters used in the system (number of cutpoints, required prevalences, degrees of patterns, number of pattern who cover each observation, etc.) were established through the usual calibration method of *LAD* by experimentally identifying the parameter values providing the highest accuracy levels.

2.2.1. Support set selection The *LAD* method starts with a pre-processing procedure for the selection of a significant support set of genes, on which the proposed prognostic system will be constructed. Since these systems are expected to have high accuracy, we restricted our study only to those 13,387 genes whose log-ratio measurements of fluorescence intensities are known for every single patient, i.e. we have eliminated those genes which include missing data. Further, we have retained from this set only those 10,665 genes which have no unidentified sequence tags (EST’s).

This set of 10,665 genes was subsequently “filtered” using the support set construction method of [5] and produced a support set consisting of only 16 genes. This set was used subsequently as the set of independent variables of a *LAD* system, and is presented in Table 1.

| # | Accession # | Gene Description |
|----|-------------|--|
| 1 | NM_020123 | Homo sapiens endomembrane protein emp70 precursor isolog (LOC56889), mRNA. |
| 2 | NM_004994 | matrix metalloproteinase 9 (gelatinase B, 92kD gelatinase, 92kD type IV collagenase) |
| 3 | NM_003748 | aldehyde dehydrogenase 4 (glutamate gamma-semialdehyde dehydrogenase; pyrroline-5-carboxylate dehydrogenase) |
| 4 | NM_016448 | L2DTL protein |
| 5 | NM_013306 | Homo sapiens clone iota unknown protein (HSAF001435), mRNA. |
| 6 | NM_014675 | KIAA0445 gene product |
| 7 | NM_018688 | bridging integrator-3 |
| 8 | NM_005243 | Ewing sarcoma breakpoint region 1 |
| 9 | NM_000221 | ketoheokinase (fructokinase) |
| 10 | NM_005744 | ariadne (Drosophila) homolog, ubiquitin-conjugating enzyme E2-binding protein, 1 |
| 11 | NM_014003 | pre-mRNA splicing factor similar to S. cerevisiae Prp16 |
| 12 | AL049689 | hypothetical protein similar to tenascin-R |
| 13 | NM_020974 | Homo sapiens CEGP1 protein (CEGP1), mRNA. |
| 14 | AF131819 | Homo sapiens cDNA: FLJ21635 fis, clone COL08233, highly similar to AF131819 Homo sapiens clone 24838 mRNA sequence |
| 15 | AL137268 | KIAA0759 protein |
| 16 | AL117502 | Homo sapiens mRNA; cDNA DKFZp434D0935 (from clone DKFZp434D0935) |

Table 1. Support Set of 16 Genes.

2.2.2. Binarization We have used a simple binarization technique which introduces two cutpoints into the range of fluorescence intensities of each gene, dividing it into three zones (low, medium, and high). The cutpoints for each particular gene were defined in such a way that the number of cases in the training set having low, medium, or high expression levels for that gene should be approximately equal.

2.2.3. Pattern generation In order to assure the high reliability of the patterns used in the model, we have restricted our search to patterns of prevalence at least 35%. Further, in order to maximize the explanatory power of the patterns detected, we have restricted our search to patterns of degree at most 3 (i.e., involving at most 3 genes). In this way, we have identified the 39 positive and 93 negative patterns shown in Tables A1 and A2 (see [2]); two of these patterns are shown in Table 2.

| Pattern | Names of genes in the support set and pattern defining inequalities | | | | | | | | | | | | | | Prevalence (%) | | | | | |
|---------|---|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|----------|-----------|----------|----------------|----------|----------|----------|----------|----------|
| | NM_020123 | NM_004994 | NM_003748 | NM_016448 | NM_013306 | NM_014675 | NM_018688 | NM_005243 | NM_000221 | NM_005744 | NM_014003 | AL049689 | NM_020974 | AF131819 | AL137268 | AL117502 | Positive | Negative | Positive | Negative |
| P9 | > -0.063 | | ≤ -0.0265 | | | | ≤ 0.0365 | | | | | | | | | | 41.2 | 0.0 | 41.7 | 0.0 |
| N2 | | | | | | | | > -0.028 | > 0.0455 | | | | > -0.546 | | | | 0.0 | 52.3 | 0.0 | 28.6 |

Table 2. Examples of Patterns (Extracted from Tables A1 and A2).

Each row in Tables A1, A2, as well as in Table 2 represents a pattern. For example, the row labeled P9 in Table 2 represents a positive pattern, defined by the three conditions “Gene NM_020123 > - 0.063, Gene NM_003748 ≤ - 0.0265, and Gene NM_018688 ≤ 0.0365”. This pattern is verified by 41.2% of the positive cases and none of the negative cases in the training set, as well as by 41.7% of the positive cases and none of the negative cases in the test set.

2.2.4. Prognosis The availability of the pandect makes it possible to *classify* new (i.e., not yet seen) observations as being positive or negative. Prognosis is one of the most important applications of LAD to biomedical problems. The most direct way to apply LAD to prognostic problems is to examine which patterns are displayed by a new case. If the case displays only positive patterns, then it is assigned a poor prognosis. Similarly, if it displays only negative patterns, then it is assigned a good prognosis. If the case does not display any pattern, then no prognosis can be assigned to it; it should be remarked that this situation is extremely rare, and did not occur at all in this study. Finally, if a case displays both positive and negative patterns, then a simple weighting procedure is applied to determine whether the positive or the negative patterns are predominant. The weighting procedure consists simply in comparing the proportion of the displayed positive patterns in the set of all positive patterns contained in the model (pandect), with the analogous proportion of negative patterns. For instance, a patient displaying 5 of the 39 patterns (i.e., 12.8%) in the positive pandect and 3 of the 93 patterns (i.e., 3.2%) in the negative pandect is assigned a poor prognosis.

2.2.5. Calibration The quality of the prognosis given by the pandect is a consequence of the choice of several *control parameters*. The major control parameters include: the number of cutpoints per gene, upper bounds on the size of support sets, pattern degrees, lower bounds on pattern prevalence. The control parameters define uniquely the pandect. The best values of the control parameters are

determined iteratively by assigning some values to them, constructing the associated pandect, verifying the correctness of its predictions, reassigning the values, and continuing this sequence of steps until arriving to a pandect with highly accurate predictions. The verification process is based on well-known statistical cross-validation techniques.

The most frequently used cross-validation techniques are those of *leave-one-out* (or *jackknifing*) method, and *k-folding* (e.g. [18]). All the cross-validation techniques are carried out within the training set, i.e., they do not involve any observation in the test set. In *leave-one-out*, one of the cases is taken as *verification* set, the pandect is built on the remaining cases (the *learning* set), and its prognosis is checked on the unique case in the verification set; this experiment is repeated for each case in the training set. In *k-folding*, the training set is partitioned randomly into *k* (e.g., 2, 5 or 10) subsets; one of these subsets is then selected as the verification set, the pandect is constructed on the remainder of the training set (viewed as the learning set), and the prognosis of the pandect is checked on the verification set. This experiment is repeated *k* times, for each of the *k* possible selections of the verification set.

2.2.6. Validation of the *LAD* results can be carried out in two ways. First, the predictions of the pandect built on the training set has to be checked on the test set. This is the most frequently used validation method. In order to increase the reliability of the proposed pandect, an additional validation procedure can be applied. In this second validation procedure, a new dataset is created, consisting of all the observations in the original training and test sets. The second validation consists now in the application of the usual cross-validation techniques (*k-folding* and/or *leave-one-out*) to this augmented dataset, using the parameters found at the calibration stage.

3. RESULTS

3.1. Prognostic System

One of the main reasons for the construction of the pandect was to create prognostic systems capable of distinguishing patients with poor prognosis from those with good prognosis. It turns out that the pandect of 39 positive and 93 negative patterns, constructed in this study, provides highly accurate prognoses.

The classification provided by the pandect for the 34 positive and the 44 negative observations in the training set makes no errors (accuracy = 100%). More significantly, on the 19-element test set, the system makes only one error, and classifies correctly all the other cases; thus, the system's accuracy is 94.7%. The only error is of type 2, and it is due to the incorrect classification of the negative sample 119. The supplementary validation tests based on an additional series of 50 five-folding experiments on the combined dataset of 97 cases shows an average accuracy of 85.8%.

3.2. Significant Biomarkers

The analysis of genes appearing in the description of the patterns in Tables A1 and A2 shows that the expression level of gene *NM_003748 aldehyde dehydrogenase 4 (glutamate gamma-semialdehyde dehydrogenase; pyrroline-5-carboxylate dehydrogenase)* which appears in a large number (51% of the positive, and 33% of the negative) patterns, has a high significance in the prognosis signature.

Based on the frequency of inclusion of genes in the positive patterns, it can be seen that the following five genes play a significant role in determining a poor prognosis: *NM_003748 aldehyde dehydrogenase 4 (glutamate gamma-semialdehyde dehydrogenase; pyrroline-5-carboxylate dehydrogenase)*, *NM_014675 KIAA0445 gene product*, *NM_013306 Homo sapiens clone iota unknown protein (HSAF001435), mRNA*, *NM_020974 Homo sapiens CEGP1 protein (CEGP1), mRNA*, and *NM_005744 ariadne (Drosophila) homolog, ubiquitin-conjugating enzyme E2-binding protein, 1*. Similarly, based on the frequency of inclusion of genes in the negative patterns, it can be seen that the following four genes play a significant role in determining a good prognosis: *NM_016448 L2DTLprotein*, *NM_020974 Homo sapiens CEGP1 protein (CEGP1), mRNA*, *NM_003748 aldehyde*

dehydrogenase 4 (glutamate gamma-semialdehyde dehydrogenase; pyrroline-5-carboxylate dehydrogenase), and *NM_013306 Homo sapiens clone iota unknown protein (HSAF001435), mRNA*.

It is important to remark that the genes *NM_003748* and *NM_013306* appear to be significant for both the poor and the good prognosis signatures.

3.3. Contributors and Blockers

A gene with the property that an increase in the intensity level of its expression, while the expression levels of the other genes remain unchanged, can sometimes worsen the prognosis, but can never improve it, will be called in this paper a *contributor*. Similarly, a gene with the property that a decrease in the intensity level of its expression, while the expression levels of the other genes remain unchanged, can sometimes improve the prognosis, but can never worsen it, will be called in this paper a *blocker*. Clearly, not every gene is a contributor or a blocker.

The pandect can identify contributors and blockers. In order to extract this information, we shall have to complete the description of the pandect as shown in Tables B1 and B2 (see [2]) by filling in some of the empty entries. Each row of the table corresponds to a pattern. Those limitations on the expression levels of genes which define the pattern, appear as bounding constraints (\leq or $>$) in the table. However, beside these constraints defining a pattern, the expression levels of some of the other genes may also be restricted for all the cases covered by the pattern.

For illustration, pattern *P5* in Table A1 is defined by the following limitations imposed on the intensity levels of expressions of three genes: “*NM_003748* \leq -0.0265, *NM_013306* \leq -0.008, *NM_020974* \leq 0.086”. However, it can be seen that the cases displaying this pattern satisfy the additional limitations: “*NM_014675* \leq 0.0535, *NM_005744* $>$ - 0.043”. These additional limitations have to be introduced now in an updated form of the pandect. Note, that the pandect did not change by this new representation, only its description is more complete. We shall call this the *explicit form* of the pandect.

The explicit form of the pandect allows the straightforward identification of the contributors and blockers. Indeed, a gene without upper bounding conditions in the explicit form of the positive pandect (i.e., a gene whose column in the table representing the explicit form of the pandect does not contain any inequality of the \leq type) is a contributor. Similarly, a gene without lower bounding conditions in the explicit form of the positive pandect (i.e., a gene whose column in the table representing the explicit form of the pandect does not contain any inequality of the $>$ type) is a blocker.

The explicit form of the pandect (Tables B1 and B2) identifies the 6 contributors (*NM_020123*, *NM_004994*, *NM_016448*, *NM_005744*, *NM_014003*, *AF131819*) and the 10 blockers (*NM_003748*, *NM_013306*, *NM_014675*, *NM_018688*, *NM_005243*, *NM_000221*, *AL049689*, *NM_020974*, *AL137268*, *AL117502*) contained in the support set. Further blockers and contributors can be identified by the examination of additional support sets. It is important to note that the fact that each of the genes in this support set has a consistent behavior (being either a contributor or a blocker) is highly unusual, since most datasets contain very few attributes with a consistent behavior.

3.4. Special Classes

3.4.1. Prominent Positive Class Using the pattern-based representation of cases described in section 2.1.5., we have examined the class of all those positive cases in the training set, which satisfy at least a fixed proportion of the 39 positive patterns associated to the support set. By requiring this proportion to be 48%, we have identified a particularly interesting class with consistent characteristics. This class *P* contains the 13 cases # 48, 50, 54, 61, 65, 66, 67, 69, 70, 71, 74, 75, 76.

The 12 cases in *P* satisfy a number of interesting common characteristics:

(i) *Multiple Coverage by Patterns*: the average case in *P* satisfies 65.59% of the positive patterns, while this average for the positive cases not in *P* is 28.80%;

(ii) *Predictability*: in each validation experiment of the prognostic system by the leave-one-out method, the prognosis of every single observation in P was correct; the 100% sensitivity of the prognostic system on the set P is much higher than its 80.95% sensitivity on the set of positive cases not contained in P;

(iii) *Distinctive Gene Expression Ranges*: the class P can be described (Table C1, see [2]) in terms of the expression levels of the intensities of the genes in the support set; these conditions are characteristic for the prominent class P, since no other case in the training or test set satisfies them;

(iv) *Statistical Distinctions of Clinical Features*: the contrast between the average measurements of clinical parameters of the positive observations in the class P and of those outside of the class P can be seen in Table D1 (see [2]). It is clear that that some of these measurements show substantial differences; in particular, the average values of the diameter, estrogen receptor positive, and lymphocytic infiltrate in the class P are outside of the 95% confidence intervals of the corresponding parameters for the positive observations not contained in P.

3.4.2. Prominent Negative Class In perfect analogy with the positive case, we have identified a prominent class N of 17 negative cases, distinguished from the other negative cases by the fact that each of them satisfies at least 43% of the negative patterns associated to the support set. The class N consists of the cases # 1, 6, 7, 13, 14, 17, 18, 19, 21, 22, 27, 28, 29, 31, 33, 36, 42. Similarly to P, the class N has several distinguishing features:

(i) *Multiple Coverage by Patterns*: the average case in N satisfies 67.99% of the negative patterns, while this average for the negative cases not in N is 20.36%;

(ii) *Predictability*: in each validation experiment of the prognostic system by the leave-one-out method, the prognosis of every single observation in N was correct; the 100% specificity of the prognostic system on the set N is much higher than its 77.77% specificity on the set of negative cases not in N;

(iii) *Distinctive Gene Expression Ranges*: the class N can be described (Table C2, see [2]) in terms of the expression levels of the intensities of the genes in the support set; these conditions are characteristic for the prominent class N, since no other case in the training or test set satisfies them;

(iv) *Statistical Distinctions of Clinical Features*: the contrast between the average measurements of clinical parameters of the positive observations in the class N and of those outside of the class N can be seen in Table D2 (see [2]). While some differences can be noticed between the average measurements in the two classes, and none of these differences contradicts the notion that the cases in N are “more negative” than those outside of N, it is not clear whether these measurements – with the possible exception of Grade – show very substantial differences.

4. DISSCUSSION

4.1. Comparative Accuracy

On the **training set** of 34 positive and 44 negative cases, the model of [17] misclassifies 12 positive and 3 negative cases. The proposed pandect-based prognostic system identifies correctly 100% of the cases in the training set. On the 19-element **test set**, the model of [17] misclassifies 2 cases, while the *LAD* system misclassifies 1.

We are not aware whether the performance of the model presented in [17] has been subjected to cross-validation by *k*-folding or leave-one-out experiments, and can therefore not compare them with the cross-validation results of *LAD* reported in Section “Results”.

Table 3 below gives a clear indication of the performance and robustness of the proposed *LAD*-based prognostic system.

| | <i>Training set (78 cases)</i> | | <i>Test set (19 cases)</i> | <i>Entire dataset (78+19 cases)</i> |
|--------------------------|--------------------------------|-------------------------|------------------------------|-------------------------------------|
| | <i>Direct Classification</i> | <i>Cross-Validation</i> | <i>Direct Classification</i> | <i>Cross-Validation</i> |
| <i>Classifier ([17])</i> | 65 (=83.3%) | Not Reported | 17 (=89.5%) | Not Reported |
| <i>LAD Prognosis</i> | 78 (=100%) | 67 (=85.9%) | 18 (=94.7%) | 83.3 (=85.8%) |

Table 3. Number and Percentage of Correct Prognoses.

4.2. Individual vs. Collective Biomarkers

One of the important hypotheses raised by the *LAD* approach concerns the role played in an accurate diagnostic system by those genes which have the highest correlation with the outcome. In contrast with the conventional approach, *LAD* aims at going beyond the straightforward goal of identifying genes with important *individual* contributions to cancer detection (as measured by their correlation with the outcome), focusing on those genes which taken as a group, have the highest *collective* prognostic potential.

The breast cancer prognostic system developed in this study confirms the hypothesis that the most accurate prognostic systems must not necessarily include only genes with high correlations with the outcome. Indeed, the 70 biomarkers used in the van't Veer study are extracted from the pool of the 231 genes which (taken individually) are most highly correlated with the outcome. On the other hand, the 16 gene support set selected by *LAD* includes several genes whose correlation with the outcome in absolute value is very low.

It is interesting to remark that (i) the overlap between the set of 16 genes selected by *LAD* and the set of the 70 genes used by in the van't Veer study ([17]) consists of only 2 genes (*NM_003748* and *NM_020974*), (ii) the overlap between the set of 16 genes and the pool of 231 genes from which the 70 biomarkers have been extracted ([17]) consists of only 4 genes (the two mentioned above and *NM_004994* and *NM_016448*).

The high accuracy of the *LAD* model is not due to the individual role of the selected genes, but to the interactions among various genes in the “collective biomarkers” corresponding to the patterns. It should be added that the average of the absolute values of the Pearson correlations of these 132 positive and negative collective biomarkers with the outcome (0.48) is substantially higher than that of the genes included in the support set (0.36).

4.3. Contrast between training and test sets

The result shown above for the accuracy of the prognostic system on the test set (94.7%) is “too good to be true”, since its accuracy in the cross-validation experiments on the training set was substantially smaller (85.9%). In fact, this discrepancy is perhaps due to the fact that the training and the test sets in this dataset have different characteristics. Further, the closeness of the cross-validation experiments on the training set (85.9%) and on the entire dataset (85.8%) reinforces the idea of the presence of a discrepancy between the training and the test sets.

The same idea is reinforced by the fact that it is easy to identify genes which – even when considered one-at-a-time --are capable to distinguish completely the two sets. For example the cutoff value -0.4625 of the gene *Contig42177_RC* separates perfectly the training and the test sets. The same applies to the cutoff value -0.3355 of the gene *AB020720*, as well for the cutoff value -0.145 of the gene *NM_018613*. Clearly, this separation is even stronger than that given by the 100% accurate Fisher discriminant $205.60 * Contig42177 + 8.57 * AB020720 + 18.47 * NM_018613$, whose values for the cases in the training set exceed -114.07, while for those in the test set they are below the same threshold.

In addition to these genes, we have also identified by *LAD* a set of 13 other genes (*Contig54839_RC*, *NM_012075*, *Contig53315_RC*, *NM_020552*, *NM_004649*, *NM_006256*, *NM_005649*, *NM_005760*, *Contig40513_RC*, *NM_014895*, *Contig67163*, *NM_000121*, *Contig48877_RC*), which show markedly different values on the training and the test sets. Numerous low-degree patterns (Table E, see [2]), including 67 degree 2 ones distinguish perfectly the training and the test sets.

The existence of pairs of genes which can distinguish between training and test observations, is an extremely rare situation. The existence of individual genes allowing such a distinction is obviously even more surprising. Even in datasets where the training and test samples are collected in different

laboratories the existence of such genes or pairs of genes is highly unlikely. For instance, no such separation exists for the microarray dataset *Leukemia AML-ALL* studied in [13].

Moreover, it can be seen that the genes in the support set define a "box" (i.e., an interval in the 16-dimensional real space) – described in Table F (see [2]) – includes all the 19 test cases and none of training cases, thus providing a complete separation of the two sets.

In light of the conclusions of this section, we would expect the accuracy of the *LAD*-based prognostic system applied to new cases to be predicted more accurately by its 85.9% performance in cross-validation, than by its 94.7% performance on the test set. Similar prudence should apply to the evaluation of the accuracy of any other model learned on this training set and measured on this test set.

4.4. Future Studies

4.4.1. Individualized Therapy

An important consequence of the identification of contributors and blockers is the possibility of targeting therapies in such a way that they should raise the expression levels of the intensities of some of the blockers, and/or lower those of some of the contributors. An even more attractive challenge is that of developing individualized therapies which target the particular blockers and contributors present in the specific positive and negative patterns which are "triggered" by the expression levels of an individual's genes.

4.4.2. Prognostic Index

The results presented in the subsection *Prominent Classes* (3.4.1) indicate the existence of a possibly strong correlation between prognosis accuracy and the proportion of patterns covering a case. Similarly to the index introduced in [5] for risk stratification among cardiac patients, it is to be expected that a prognostic index for breast cancer patients could also be developed.

References

- [1] Abramson, S., Alexe, G., Hammer, P.L., Kohn, J. Using Logical Analysis of Data (LAD) Based Computer Model Predicts Cell Metabolic Activity on Polymeric Substrates, *RUTCOR Research Report*, RRR 40-2002; Communication at the 29th Annual Meeting of the Society for Biomaterials, Reno, Nevada, April-May 2003.
- [2] Alexe, G., Alexe, S., Axelrod, D., Boros, E., Hammer P. L., Reiss, M., Appendix to "Combinatorial analysis of breast cancer data from gene expression microarrays", http://rutcor.rutgers.edu/~alexe/Appendix_LAD_BC.xls
- [3] Alexe, G., Alexe, S., Hammer, P.L., Kogan, A. Comprehensive vs. Comprehensible Classifiers in Logical Analysis of Data. Rutgers University, *RUTCOR Research Report*, RRR 9-2002; *DIMACS Technical Report 2002-49*; *Annals of Operations Research* (in print).
- [4] G. Alexe, S. Alexe, P.L. Hammer, L. Liotta, E. Petricoin, M. Reiss. Logical Analysis of the Proteomic Ovarian Cancer Dataset. *RUTCOR Technical Report*, RTR 2-2002 (<http://rutcor.rutgers.edu/~rrr/rtr/2-2002.pdf>).
- [5] Alexe, G., Alexe, S., Hammer, P. L., Vizvari, B. Pattern-Based Feature Selection in Genomics and Proteomics. Rutgers University, *RUTCOR Research Report*, RRR 7-2003.
- [6] Alexe, G., Hammer, P. L. Spanned Patterns in Logical Analysis of Data. Rutgers University, *RUTCOR Research Report*, RRR 15-2002, *Annals of Operations Research* (2003), (in print).
- [7] Alexe S., Blackstone E., Hammer, P. L., Ishwaran, H., Lauer, M. S., Pothier Snader, C. E. Coronary Risk Prediction by Logical Analysis of Data. *Annals of Operations Research*, 119 (2003), 15-42.
- [8] Alexe, S., Hammer, P.L. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. *RUTCOR Research Report*, RRR 59-2002, *Annals of Operations Research* (2003), (in print).
- [9] Alexe, S., Hammer, P.L., Kogan, A., Lejeune, M.A. A Non-Recursive Regression Model For Country Risk Rating, *RUTCOR Research Report*, RRR 9-2003.

- [10] Boros E., Hammer P.L., Ibaraki T., Kogan A. Logical Analysis of Numerical Data. *Mathematical Programming* 79 (1997), 163-190.
- [11] Boros, E., Hammer, P.L., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I. An Implementation of Logical Analysis of Data. *IEEE Transactions on Knowledge and Data Engineering*, 12 (2) (2000), 292-306.
- [12] Crama, Y., Hammer, P.L., Ibaraki, T. Cause-Effect Relationships and Partially Defined Boolean Functions. *Annals of Operations Research* 16 (1988), 299-326.
- [13] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield, C. D., Lander E. S. Molecular Classification of Cancer; Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286 (5439) (1999), 531-537.
- [14] Hammer, A., Hammer, P.L., Muchnik, I. Logical Analysis of Chinese Productivity Patterns, *Annals of Operations Research* 87, 1999, 165-176.
- [15] Eckstein, J., Hammer, P.L., Liu, Y., Nediak, M., Simeone, B. The Maximum Box Problem and its Application to Data Analysis. *Computational Optimization and Applications* (in print).
- [16] Lauer, M.S., Alexe, S., Snader, C.E.P., Blackstone, E., Ishwaran, H., Hammer, P. L. Use of the "Logical Analysis of Data" Method for Assessing Long-Term Mortality Risk After Exercise Electrocardiography. *Circulation*, 106 (2002), pp. 685-690.
- [17] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* 415 (2002), 530-535.
- [18] S-Plus 6 for Windows: Guide to Statistics, Vol. 1 and Vol. 2, Insightful Corporation, Seattle, WA, 2001.