
Exploring Microarray Data with Correspondence Analysis

Stanislav Busygin and Panos M. Pardalos

Industrial and Systems Engineering Department
University of Florida
303 Weil Hall, Gainesville, FL 32611
{busygin,pardalos}@ufl.edu

Summary. Due to the rapid development of DNA microarray chips it has become possible to discover and predict genetic patterns relevant for various diseases on the basis of exploration of massive data sets provided by DNA microarray probes. A number of data mining techniques have been used for such exploration to achieve the desirable results. However, high dimensionality and uncertain accuracy of microarray datasets remain the major obstacles in revealing the most crucial genetic factors determining a particular disease. This chapter describes a microarray data processing technique based on the correspondence analysis that helps to handle this issue.

1 Introduction

The importance of data analysis in life sciences is steadily increasing. Up to recently, biology was a descriptive science providing relatively small amount of numerical data. However, nowadays it has become one of the main applications of data mining techniques operating on massive data sets. This transformation can be particularly attributed to two recent advances which are complementary to each other. First, the Human Genome Project and some other genome-sequencing undertakings have been successfully accomplished. They have provided the DNA sequences of the human genome and the genomes of a number of other species having various biological characteristics. Second, revolutionary new tools able to monitor quantitative data on the genome-wide scale have appeared. Among them, there are the *DNA microarrays* widely used at the present time. These devices measure gene expression levels of thousands of genes simultaneously, allowing researchers to observe how the genes act in different types of cells and under various conditions.

As a consequence of this progress, the traditional approach of studying one particular gene per experiment has been changed. Now it is possible to investigate not only how a gene behaves itself, but also how it *interacts* with other genes and which gene expression patterns are formed. It is natural to expect

that on the basis of microarray data, the genes characterizing certain medical phenomena (such as diseases) can be detected and classified. Especially, such a study is crucial for understanding *genetic diseases* caused by a mutation in a gene or a set of genes. They make the mutant genes inappropriately expressed or even not expressed at all. For example, it is known that cancer can be caused by inactivation, deletion or, on the contrary, by constitutive activity of *p53 tumor suppressor gene*. Furthermore, some genetic diseases have subtypes that are indistinguishable clinically but differ from each other in the underlying genetic mechanism. Most likely, it would imply that these subtypes require different methods of treatment. However, unless a sophisticated diagnostic technique is available, it would be impossible to properly make the right choice. One illustrative example of such a situation considered in this chapter is discriminating *acute lymphoblastic leukemia* (ALL) versus *acute myeloid leukemia* (AML).

However, the analysis of microarray data is not an easy task. High dimensionality of the data, poor accuracy of microarray probes, and practical difficulties with taking the probes (the procedure might be very painful for alive patients while the gene expression levels rapidly degrade in dead tissue) hinder the success of microarray technology. Hence, the microarray datasets must be processed by a sophisticated data mining technique applicable in the case of high-dimensional data and still able to refine particular data values known to be critically inaccurate.

Generally, data mining techniques may be divided into three major classes that sometimes overlap: statistical analysis, clustering, and dimensionality reduction (projection methods). The statistical analysis for microarray data usually consists in calculating *fold change* of particular genes across different groups of samples and applying classical statistic tests such as *t*-test, ANOVA, Wilcoxon test, etc. These techniques are appropriate when a proper separation of samples into classes is known, the number of outliers in each class is insignificant, and the data may be assumed to have certain statistical properties (e.g., normal distribution). While the normality assumption is believed to be feasible for microarray data [4], the other conditions are harder to guarantee, taking into account the issues with accuracy of microarray data mentioned above. Furthermore, the statistical analysis cannot reveal more general patterns in the data rather than up- or downregulation of single genes.

Clustering techniques can be divided into *supervised* and *unsupervised* learning. Supervised learning techniques are also called sometimes *classification* methods. They take predetermined classes of objects as input and aim at deriving characteristics (features) common for samples of a class and discriminating them against samples of other classes. Examples of supervised learning techniques include linear discriminant analysis, classification and regression trees, support vector machines, etc. Clearly, supervised learning requires a set of *training samples* whose separation into different classes is known beforehand. On the contrary, unsupervised clustering techniques do not require such a training set; they build classes (clusters) of samples starting from scratch.

Examples of clustering techniques are hierarchical clustering, k-means clustering, self-organizing maps (SOM), etc. Common drawbacks of these methods are significant dependence of the results on initialization of the clustering and the absence of a clear mathematical criterion to judge quality of the results. (i.e., a universal objective function whose optimal value would signify best clustering in all instances does not exist). Furthermore, it has been proved in [8] that there is no clustering algorithm simultaneously satisfying three simple properties one might expect to be required: *scale-invariance* (i.e. multiplying all distances by the same positive number should not change the result), *richness* (all partitions should be achievable), and *consistency* (decreasing the distances within the clusters with increasing the distances between the clusters should not change the result).

The dimensionality reduction methods do not aim at delivering strict categorization of data into classes or separation of relevant versus non-relevant features. They rather produce a low-dimensional projection of an originally high-dimensional data set. As soon as such a projection is presented in the form of biplot or 3D diagram, there is the opportunity for a researcher of the data domain to eyeball the picture and gain an understanding of the crucial data patterns. Clearly, there are at least two advantages comparing to strict categorization of the data. First, when the data dimensionality is small, the human eye becomes an analytic tool of remarkable power able to grasp complex data patterns undetectable by any statistical methods generally aimed at simple linear relations. Second, optimization of projecting high-dimensional spaces onto low-dimensional subspaces is nicely supported by extensive theoretical background of linear algebra. The core of projection methods is *singular value decomposition* (SVD), which can provide the subspace of any desirable dimension preserving the maximum possible similarity between the original data set and its projection onto the subspace. Another important point here is that the SVD procedure is computationally efficient (i.e. it can be performed in a short polynomial time, especially if only few dominating singular vectors are sought). This compares favorably to many iterative clustering procedures. For instance, the convergence of SOM cannot even be guaranteed without gradually decreasing the learning rate parameter with each iteration. Hence the projection techniques are also attractive from the computational complexity viewpoint. Lastly, the projection techniques do not depend on any parameters that should be specified by the user before the algorithm is applied. This potentially makes them more appealing for biological researchers typically not familiar in detail with the data mining algorithms. We refer the reader to [5] for detailed introduction of relevant algebra and SVD algorithms.

In this chapter we describe one specific dimensionality reduction technique called *correspondence analysis* (CA), and consider its application to microarray data. The effectiveness of CA is illustrated by discovering AML- and ALL-relevant genes from a well-known microarray dataset published by Golub et al [6].

2 Correspondence Analysis

2.1 Basic Algorithm

Like other dimensionality reduction techniques, correspondence analysis is an *exploratory data analysis* technique providing a view of the data set as a whole. The main advantage of correspondence analysis over other dimensionality reduction techniques is that it allows for simultaneous observation of data samples (usually given by columns of the data matrix) and data points (correspondingly, represented by rows of the data matrix) in *one* low-dimensional space. This becomes possible due to the bidirectional nature of correspondence analysis, investigating not only relations within the set of samples and the set of data points, but also cross-relations between elements of these two sets. The only restriction of this technique is that all data values must be nonnegative.

Thus, correspondence analysis maps all samples and data points of a data set onto one low-dimensional space, which can be visualized as a biplot (2-D) or 3-D diagram. Each axis of this diagram tends to reveal a profound characterization of the data set, and samples/data points having high similarity with respect to this characterization have similar coordinates on it. Like in other dimensionality reduction techniques, the construction of the low-dimensional space is performed by means of singular value decomposition (SVD). However, in case of the correspondence analysis, SVD is not applied directly to the data matrix, but is used after its specific *correspondence* matrix is constructed. We refer the reader to the existing literature (e.g., [7]) to review theoretical background of the method and related algebraic proofs. Here we describe the algorithm of correspondence analysis and its generalization in case when some data entries are missing or cannot be trusted.

A data set is normally given as a rectangular matrix $A = (a_{ij})_{m \times n}$ of n samples (columns) and m data points (rows). In the case of microarray data, rows represent genes and the value a_{ij} shows the expression level of gene i in sample j . For the sake of simplicity, we assume further on that $m > n$. However, this does not restrict the generality of the discussed technique, since both the columns and the rows of the data matrix are to be treated in a unified way, and it is always possible to work with the transposed matrix without making any changes in the algorithm. So, to perform correspondence analysis, we first construct the *correspondence* matrix $P = (p_{ij})_{m \times n}$ by computing

$$p_{ij} = a_{ij}/a_{++}, \quad (1)$$

where

$$a_{++} = \sum_{i=1}^m \sum_{j=1}^n a_{ij} \quad (2)$$

is called the *grand total* of A . The correspondence matrix is somewhat analogous to a two-dimensional probability distribution table, whose entries sum

up to 1. We also compute *masses* of rows and columns (having the analogy with the marginal probability densities):

$$r_i = a_{i+}/a_{++}, \quad (3)$$

$$c_j = a_{+j}/a_{++}, \quad (4)$$

where

$$a_{i+} = \sum_{j=1}^n a_{ij}, \quad (5)$$

$$a_{+j} = \sum_{i=1}^m a_{ij}. \quad (6)$$

Then the matrix $S = (s_{ij})_{m \times n}$, to which SVD is applied, is formed:

$$s_{ij} = (p_{ij} - r_i c_j) / \sqrt{r_i c_j}. \quad (7)$$

The SVD of $S = UAV^T$ represents the provided matrix as the product of three specific matrices. Columns of the matrix $U = (u_{ij})_{m \times n}$ are orthonormal vectors spanning the columns of S , columns of the matrix $V = (v_{ij})_{n \times n}$ are also orthonormal vectors but they span the rows of S , and finally $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is a diagonal matrix of nonnegative *singular values* of S having a nondecreasing order: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$. It can be shown algebraically that the optimal low-dimensional subspace to project the columns of S onto, with the minimum possible information loss, is formed by a desired number of first columns U . Similarly, the optimal low-dimensional subspace for plotting the rows of S is formed by the same number of first columns of V . Furthermore, due to specific properties of the matrix S , the columns and rows of the original data set matrix A may be represented in one low-dimensional space of dimensionality $K < n$ as follows:

$$f_{ik} = \lambda_k u_{ik} / \sqrt{r_i}, \quad k = 1, 2, \dots, K, \quad (8)$$

gives the k -th coordinate of row i , and

$$g_{jk} = \lambda_k v_{jk} / \sqrt{c_j}, \quad k = 1, 2, \dots, K, \quad (9)$$

gives the k -th coordinate of column j in the new space. Obviously, we select $K = 2$ if we want to obtain a biplot and $K = 3$ if we want to obtain a 3-D diagram of the analyzed data set.

2.2 Treatment of missing values

Correspondence analysis allows for an easy and natural treatment of missing data values. We just need to look at the procedure backward and answer

the question: if f and g were the positions of rows and columns on the low-dimensional plot, what value of a data entry a_{ij} would minimize the information loss incurred due to the dimensionality reduction with respect to the constructed low-dimensional representation? It is necessary to mention here that the correspondence analysis algorithm constructing the low-dimensional space actually solves the following least-squares problem [7]:

$$\min \sum_{i=1}^m \sum_{j=1}^n (a_{ij} - \hat{a}_{ij})^2 / (a_{i+} a_{+j}), \quad (10)$$

where $\hat{A} = (\hat{a}_{ij})_{m \times n}$ is the sought low-dimensional approximation of the data that can be expressed as

$$\hat{a}_{ij} = (a_{i+} a_{+j} / a_{++}) \left(1 + \sum_{k=1}^K f_{ik} g_{jk} / \sqrt{\lambda_k} \right). \quad (11)$$

So, the relation (11) gives the best guess for the data entry a_{ij} provided that we already have the low-dimensional coordinates f and g , and the singular values λ . From here we can infer an iterative *E-M algorithm* performing simultaneously construction of the low-dimensional plot of the data and approximation of missing data entries (the latter is called *imputing* the values) [7].

1. Make some initial guesses for the missing data entries.
2. Perform the K -dimensional correspondence analysis as specified by the formulas (1)-(9) (the M-step, or *maximization* step of the E-M algorithm).
3. Obtain new estimations for the imputing data entries by (11) (the E-step, or *expectation* step of the algorithm).
4. If the new estimations are close enough to the previous estimations, STOP. Otherwise repeat from Step 2 with the new estimations.

The initial guesses for the imputing data entries for Step 1 of the algorithm should be made such that $p_{ij} = r_i c_j$ for these entries [7]. This condition is equivalent to the equalities

$$a_{ij} = a_{i+} a_{+j} / a_{++} \quad (12)$$

for the missing data entries (i, j) . To find the a_{ij} values satisfying (12), we employ a simple iterative algorithm:

1. Initialize all missing data entries with 0.
2. Compute a_{++} and all a_{i+} , $i = 1, 2, \dots, m$ and a_{+j} , $j = 1, 2, \dots, n$ by (5).
3. Compute new values a_{ij} for the missing data entries by (12). If all the new values are close enough to the previous values, STOP. Otherwise repeat from Step 2.

The E-M algorithm is known to converge properly in “well-behaved” situations (for example, no row or column should be entirely missing). This condition is plausible for most microarray experiments.

3 Test Framework

We applied correspondence analysis to a well-researched microarray data set containing samples from patients diagnosed with ALL and AML diseases [6]. It has been the subject of a variety of research papers, e.g. [1, 2, 10, 11]. This data set was also used in the CAMDA data contest [3]. Our primary concern was to try to designate genes whose expression levels significantly correlate with one of the diseases. It is natural to assume that those genes may be responsible to the development of particular conditions causing one of the considered leukemia variations. The paper [6] pursues a similar goal, but the authors used a statistical analysis algorithm followed by SOM clustering. The data set was divided into two parts – the training set (27 ALL, 11 AML samples) and the test set (20 ALL, 14 AML samples) – as the authors employed a learning algorithm. We considered it without any division since correspondence analysis does not require training. Hence there were 72 samples, 47 of which are ALL and 25 are AML. All the samples were obtained with Affymetrix GeneChipTM microarray technology and contained 7129 data points. First 59 data points were miscellaneous Affymetrix control values and were removed from the data set, the rest 7070 data points were human genes. Affymetrix GeneChipTM data values represent the difference between perfect match (PM) and mismatch (MM) probes that is expected to be significant. Usually when such a value is below 20, or even negative, it is not considered reliable. Hence we regarded all the data entries that are below 20 in the data set to be missing. Furthermore, genes having more than half of the missing entries were removed from the data set since the imputing can also not be reliable in this case. The residual 4902 data points were used in the analysis.

4 Computational Results

Fig. 1 shows the biplot obtained. It becomes immediately clear from the visual inspection that the first principal axis (horizontal) discriminates ALL samples from AML samples. Now, we may regard the genes having most positive coordinates on this axis signifying the ALL condition, while those genes having most negative coordinates there signify the AML condition. Similarly to [6], we listed top 25 ALL genes and top 25 AML genes with respect to our analysis. They are presented in Tables 1 and 2.

To validate the obtained top gene sets, we tried to estimate their relevance by observing the references made to these genes in MEDLINE articles. Each article in the MEDLINE database is categorized by the Medical Subject Headings (MeSH) [13]. We employed the same approach as the High-Density Array Pattern Interpreter (HAPI) of the University of California, San Diego [9, 12], and simply observed into which MeSH categories the articles mentioning the found genes fall predominantly. The HAPI web service reports the number of terms from each MeSH category matching genes from a provided

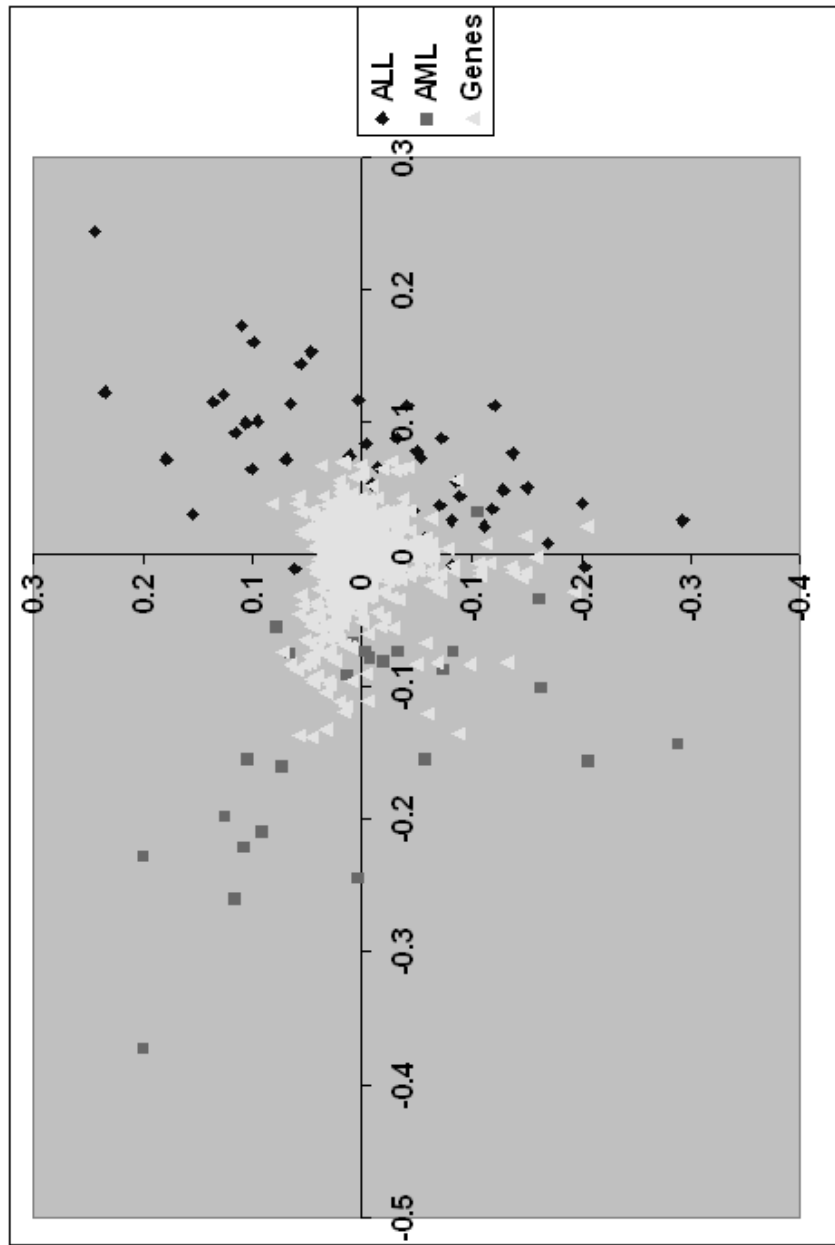


Fig. 1. Correspondence analysis biplot for the ALL vs. AML dataset

Table 1. 25 Top ALL Genes

| # | Name | Description |
|----|---------------|---|
| 1 | M89957 | IGB Immunoglobulin-associated beta (B29) |
| 2 | K01911 | NPY Neuropeptide Y |
| 3 | AF009426 | Clone 22 mRNA, alternative splice variant alpha-1 |
| 4 | D13666 s | Osteoblast specific factor 2 (OSF-2os) |
| 5 | M83233 | TCF12 Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4) |
| 6 | D87074 | KIAA0237 gene |
| 7 | X82240 rna1 | TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/lymphoma 1 |
| 8 | S50223 | HKR-T1 |
| 9 | X53586 rna1 | Integrin alpha 6 (or alpha E) protein gene extracted from Human mRNA for integrin alpha 6 |
| 10 | D88270 | GB DEF = (lambda) DNA for immunoglobulin light chain |
| 11 | M38690 | CD9 CD9 antigen |
| 12 | L33930 s | CD24 signal transducer mRNA and 3' region |
| 13 | U05259 rna1 | MB-1 gene |
| 14 | U36922 | GB DEF = Fork head domain protein (FKHR) mRNA, 3' end |
| 15 | D21262 | KIAA0035 gene, partial cds |
| 16 | M94250 | MDK Midkine (neurite growth-promoting factor 2) |
| 17 | M11722 | Terminal transferase mRNA |
| 18 | M54992 | CD72 CD72 antigen |
| 19 | D25304 | KIAA0006 gene, partial cds |
| 20 | U31384 | G protein gamma-11 subunit |
| 21 | X97267 rna1 s | LPAP gene |
| 22 | M29551 | Serine/threonine protein phosphatase 2B catalytic subunit, beta isoform |
| 23 | M92934 | CTGF Connective tissue growth factor |
| 24 | X84373 | Nuclear factor RIP140 |
| 25 | X17025 | Homolog of yeast IPP isomerase |

list. Furthermore, such a report is stored online, so the matchings found for our ALL and AML genes are available for future references [14, 15].

The report for the ALL genes shows most significant matching in such categories as “Cells” (37), “Cell Nucleus” (8), “Cells, Cultured” (10), “Hemic and Immune Systems” (16), “Immune System” (10), “Neoplasms” (12), “Neoplasms by Histologic Type” (8), “Hormones, Hormone Substitutes, and Hormone Antagonists” (8), “Enzymes, Coenzymes, and Enzyme Inhibitors” (30), “Enzymes” (30), “Hydrolases” (14), “Esterases” (10), “Transferases” (8), “Amino Acids, Peptides, and Proteins” (104), “Proteins” (90), “DNA-Binding Proteins” (8), “Glycoproteins” (12), “Membrane Glycoproteins” (8), “Membrane Proteins” (24), “Membrane Glycoproteins” (8), “Receptors, Cell Surface” (12), “Receptors, Immunologic” (12), “Transcription Factors” (12), “Nucleic Acids, Nucleotides, and Nucleosides” (14), “Nucleic Acids” (10), “Immunologic and Biological Factors” (78), “Biological Factors” (26), “Biolog-

Table 2. 25 Top AML Genes

| # | Name | Description |
|----|-----------------|--|
| 1 | U60644 | HU-K4 mRNA |
| 2 | U16306 | CSPG2 Chondroitin sulfate proteoglycan 2 (versican) |
| 3 | M69203 s | SCYA4 Small inducible cytokine A4 (homologous to mouse Mip-1b) |
| 4 | M33195 | Fc-epsilon-receptor gamma-chain mRNA |
| 5 | M21119 s | LYZ Lysozyme |
| 6 | D88422 | Cystatin A |
| 7 | M27891 | CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) |
| 8 | M57731 s | GRO2 GRO2 oncogene |
| 9 | M31166 | PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta |
| 10 | D83920 | FCN1 Ficolin (collagen/fibrinogen domain-containing) 1 |
| 11 | X97748 s | GB DEF = PTX3 gene promotor region |
| 12 | M23178 s | Macrophage inflammatory protein 1-alpha precursor |
| 13 | M92357 | B94 protein |
| 14 | HG2981-HT3127 s | Epican, Alt. Splice 11 |
| 15 | X04500 | IL1B Interleukin 1, beta |
| 16 | M57710 | LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) |
| 17 | U02020 | Pre-B cell enhancing factor (PBEF) mRNA |
| 18 | Y00787 s | Interleukin-8 precursor |
| 19 | M28130 rna1 s | Interleukin 8 (IL8) gene |
| 20 | K01396 | PI Protease inhibitor 1 (anti-elastase), alpha-1-antitrypsin |
| 21 | D38583 | Calgizzarin |
| 22 | J04130 s | SCYA4 Small inducible cytokine A4 (homologous to mouse Mip-1b) |
| 23 | X62320 | GRN Granulin |
| 24 | J03909 | Gamma-interferon-inducible protein IP-30 precursor |
| 25 | M14660 | GB DEF = ISG-54K gene (interferon stimulated gene) encoding a 54 kDA protein, exon 2 |

ical Markers” (18), “Antigens, Differentiation” (18), “Antigens, CD” (12), “Cytokines” (14), “Receptors, Immunologic” (12), “Investigative Techniques” (20), “Genetic Techniques” (9), “Biological Phenomena, Cell Phenomena, and Immunity” (7), “Genetics” (60), “Genes” (7), “Genetics, Biochemical” (43), “Molecular Sequence Data” (32), “Base Sequence” (13), “Sequence Homology” (13), “Physical Sciences” (11), “Chemistry” (11), and “Chemistry, Physical” (9).

The most significant matchings for the AML genes are in the categories “Nervous System” (13), “Cells” (102), “Blood Cells” (18), “Leukocytes” (15), “Cells, Cultured” (23), “Cell Line” (13), “Cytoplasm” (13), “Hemic and Immune Systems” (51), “Blood” (18), “Vertebrates” (18), “Algae and Fungi” (14), “Fungi” (14), “Organic Chemicals” (14), “Heterocyclic Compounds” (18), “Enzymes, Coenzymes, and Enzyme Inhibitors” (40), “Enzymes” (38), “Hydrolases” (24), “Carbohydrates and Hypoglycemic Agents” (46), “Carbo-

hydrates" (46), "Polysaccharides" (36), "Glycosaminoglycans" (18), "Proteoglycans" (14), "Lipids and Antilipemic Agents" (16), "Lipids" (16), "Amino Acids, Peptides, and Proteins" (384), "Proteins" (370), "Blood Proteins" (48), "Acute-Phase Proteins" (14), "Contractile Proteins" (14), "Muscle Proteins" (14), "Cytoskeletal Proteins" (28), "Microtubule Proteins" (14), "Globulins" (14), "Serum Globulins" (14), "Glycoproteins" (62), "Membrane Glycoproteins" (26), "Proteoglycans" (14), "Membrane Proteins" (72), "Membrane Glycoproteins" (26), "Receptors, Cell Surface" (28), "Receptors, Immunologic" (20), "Nerve Tissue Proteins" (24), "Scleroproteins" (14), "Extracellular Matrix Proteins" (14), "Nucleic Acids, Nucleotides, and Nucleosides" (64), "Nucleic Acids" (34), "DNA" (20), "Nucleotides" (26), "Immunologic and Biological Factors" (262), "Biological Factors" (116), "Biological Markers" (26), "Antigens, Differentiation" (26), "Antigens, CD" (18), "Chemotactic Factors" (16), "Growth Substances" (32), "Interleukins" (18), "Toxins" (18), "Immunologic Factors" (146), "Antibodies" (16), "Antigens" (42), "Antigens, Surface" (38), "Antigens, Differentiation" (26), "Antigens, CD" (18), "Cytokines" (68), "Growth Substances" (20), "Interleukins" (18), "Monokines" (22), "Receptors, Immunologic" (20), "Specialty Chemicals and Products" (14), "Chemical Actions and Uses" (16), "Diagnosis" (17), "Laboratory Techniques and Procedures" (14), "Immunologic Tests" (13), "Investigative Techniques" (63), "Genetic Techniques" (18), "Immunologic Techniques" (16), "Immunohistochemistry" (13), "Technology, Medical" (20), "Histological Techniques" (15), "Histocytochemistry" (15), "Immunohistochemistry" (13), "Biological Phenomena, Cell Phenomena, and Immunity" (33), "Cell Physiology" (18), "Genetics" (160), "Genes" (21), "Genetics, Biochemical" (97), "Gene Expression" (15), "Gene Expression Regulation" (17), "Molecular Sequence Data" (63), "Base Sequence" (35), "Sequence Homology" (20), "Sequence Homology, Nucleic Acid" (14), "Biochemical Phenomena, Metabolism, and Nutrition" (100), "Biochemical Phenomena" (88), "Molecular Sequence Data" (61), "Base Sequence" (33), "Physiological Processes" (15), "Growth and Embryonic Development" (13), "Physical Sciences" (25).

Obviously, such literature scoring can only give an indicative measure of the quality of the obtained results. Furthermore, it should be noted that the data set contains only leukemia samples and no control samples, so it provides no information about the normal state of the gene expressions in absence of the diseases. Hence, the data analysis can only discover genes differentiating the sample classes. However, the HAPI scoring suggests that correspondence analysis enhanced by the missing data imputing feature uncovered genes highly relevant to the leukemia conditions. Moreover, the obtained numbers of matchings in the relevant MeSH categories compares favorably to the matchings of 25 ALL and AML genes reported by Golub et al. [16, 17].

5 Conclusions

Correspondence analysis is able to deliver informative projections of high-dimensional microarray data onto low-dimensional spaces. Such results in the form of pictures can be obtained in absence of any prior information about classification of samples and/or data points of the data set. In contrast to many other data mining techniques, correspondence analysis is computationally efficient, does not involve any parameters that must be tuned before the algorithm is executed, and successfully handles missing/inaccurate data values as long as their number is moderate. Furthermore, the method proves to be useful in uncovering hidden relations between groups of samples and data points (genes), possibly outperforming in efficiency more complicated statistical analysis techniques. The obtained lists on genes discriminating ALL and AML conditions may be useful for oncology researchers, providing further insights about the roles of particular human genes in the development of the acute leukemia cases.

References

1. A. Ben-Dor, L. Bruhn, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–584, 2000.
2. A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data. *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*, 2001.
3. CAMDA 2001 Conference Contest Datasets.
<http://www.camda.duke.edu/camda01/datasets/>.
4. P. J. Giles and D. Kipling. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19:2254–2262, 2003.
5. G. H. Golub and C. F. Van Loan. *Matrix Computations*, 3rd ed. (Johns Hopkins Series in the Mathematical Sciences). Baltimore, MD: John Hopkins, 1996.
6. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
7. M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984.
8. J. Kleinberg. The impossibility theorem for clustering. *Proceedings of the NIPS 2002 Conference*, 2002.
9. D. R. Masys, J. B. Welsh, J. L. Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326, 2001.
10. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik. Feature selection for SVMs. *Proceedings of the NIPS 2000 Conference*, 2001.
11. E. P. Xing and R. M. Karp. CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics Discovery Note*, 1:1–9, 2001.
12. High-Density Array Pattern Interpreter (HAPI).
<http://array.ucsd.edu/hapi/>
13. National Library of Medicine – MeSH.
<http://www.nlm.nih.gov/mesh/meshhome.html>
14. Hierarchy of keywords from literature associated with the top 25 ALL genes reported by correspondence analysis.
http://132.239.155.52/HAPI/ALL25_453.HTML
15. Hierarchy of keywords from literature associated with the top 25 AML genes reported by correspondence analysis.
http://132.239.155.52/HAPI/AML25_502.HTML
16. Hierarchy of keywords from literature associated with the top 25 ALL genes reported by Golub et al.
http://132.239.155.52/HAPI/goluball_911.HTML
17. Hierarchy of keywords from literature associated with the top 25 AML genes reported by Golub et al.
http://132.239.155.52/HAPI/golubaml_161.HTML



<http://www.springer.com/978-0-387-69318-7>

Data Mining in Biomedicine

Pardalos, P.; Boginski, V.L.; Alkis, V. (Eds.)

2007, XVIII, 580 p., Hardcover

ISBN: 978-0-387-69318-7