Study of gene expressions' correlation structures in subgroups of Chronic Lymphocytic Leukemia Patients

Athina Tsanousa, Stavroula Ntoufa, Nikos Papakonstantinou, Kostas Stamatopoulos, Lefteris Angelis

PII: S1532-0464(19)30129-7
DOI: https://doi.org/10.1016/j.jbi.2019.103211
Article Number: 103211
Reference: YJBIN 103211

To appear in: Journal of Biomedical Informatics

Received Date: 23 May 2018
Revised Date: 1 May 2019
Accepted Date: 17 May 2019

Please cite this article as: Tsanousa, A., Ntoufa, S., Papakonstantinou, N., Stamatopoulos, K., Angelis, L., Study of gene expressions' correlation structures in subgroups of Chronic Lymphocytic Leukemia Patients, Journal of Biomedical Informatics (2019), doi: https://doi.org/10.1016/j.jbi.2019.103211

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Study of gene expressions' correlation structures in subgroups of Chronic Lymphocytic Leukemia Patients

Athina Tsanousa\textsuperscript{a}, Stavroula Ntoufa\textsuperscript{b}, Nikos Papakonstantinou\textsuperscript{b}, Kostas Stamatopoulos\textsuperscript{b,c}, Lefteris Angelis\textsuperscript{a}

\textsuperscript{a} Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, atsanous@csd.auth.gr, lef@csd.auth.gr

\textsuperscript{b} Institute of Applied Biosciences, Center for Research and Technology Hellas, Thessaloniki, Greece, s_ntoufa@yahoo.gr, n.papakonstantinou83@gmail.com, kostas.stamatopoulos@gmail.com

\textsuperscript{c} Hematology Department and HCT Unit, G. Papanikolaou Hospital, Exochi, 57010, Thessaloniki, Greece

Corresponding author: Lefteris Angelis, lef@csd.auth.gr, Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece
Abstract

In chronic lymphocytic leukemia (CLL) the interaction of leukemic cells with the microenvironment ultimately affects patient outcome. CLL cases can be divided in two subgroups with different clinical course based on the mutational status of the immunoglobulin heavy variable (IGHV) genes: mutated CLL (M-CLL) and unmutated CLL (U-CLL). Since in CLL, the differentiated relation of genes between the two subgroups is of greater importance than the individual gene behavior, this paper investigates the differences between the groups’ gene interactions, by comparing their correlation structures. Fisher’s test and Zou’s confidence intervals are employed to detect differences of correlation coefficients. Afterwards, networks created by the genes participating in most differences are estimated with the use of structural equation models (SEM). The analysis is enhanced with graph modeling in order to visualize the between group differences in the gene structures of the two subgroups. The applied methodology revealed stronger correlations between genes in U-CLL patients, a finding in line with related biomedical literature. Using SEM for multigroup analysis, different gene structures between the two groups of patients were confirmed. The study of correlation structures can facilitate the detection of differential gene expression profiles in CLL subgroups, with potential applications in other diseases. Comparison of correlations can be very useful in understanding the complex internal structural differences which signify the variations of a disease.

Keywords: comparing correlations, gene expression profiles, CLL, structural equation models
1. Introduction

Chronic lymphocytic leukemia (CLL) is the most common leukemia in the western hemisphere (Chiorazzi et al., 2005). The interaction of the leukemic cells with the microenvironment is pivotal, since it provides critical survival and proliferation signals, ultimately affecting patient outcome. This interaction is facilitated through multiple receptors that are expressed on the surface of the leukemic cells and recognize different substances that are present in the surrounding environment (Ghia et al., 2008).

The most important receptor for B lymphocytes including leukemic CLL B cells is the immunoglobulin which recognizes antigens with high specificity. CLL patients can be divided into two subgroups with different clinical course based on the mutational status of the immunoglobulin heavy variable (IGHV) genes. Cases with unmutated IGHV genes (unmutated CLL, U-CLL) are generally characterized by aggressive disease and require therapy, while cases with mutated IGHV genes (mutated CLL, M-CLL) are characterized by indolent clinical course and rarely require therapy (Hamblin et al., 1999; Damle et al., 1999).

Except for the immunoglobulin, CLL cells also express many other receptors in order to sense the microenvironment including many members of the Toll like receptor (TLR) family (Ghia et al., 2008). The TLR family in humans includes TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9 and TLR10. TLRs are expressed by many cells of the immune system and play a key role in the innate immunity, since they recognize structurally conserved molecules derived from pathogens (O'Neill et al., 2013). Once the pathogens break physical barriers such as the skin or intestinal tract mucosa, they are recognized by TLRs, which initiate an immune cell response, by activating signaling cascades in the cell. Once a TLR recognizes a pathogen it is activated and converts this interaction in a biochemical signal which enters the cell. The signal is transmitted in the cell by a complex network of different proteins that interact with each other in order to enter the nucleus and regulate gene expression (Kawasaki and Kawai, 2014; Akira and Takeda, 2004).

CLL cells have been found to express TLR1, TLR2, TLR4, TLR6, TLR7, TLR8, TLR9, and TLR10 (Muzio et al. 2009, Arvaniti et al., 2011). The TLRs on CLL cells can be activated by their respective ligands and induce the signaling cascade in order to activate the cells. CLL cells can undergo proliferation or cell death through apoptosis depending on the member of the family that will be activated and/or the individual case, since M-CLL vs U-CLL cases are differentially affected by specific TLRs (Ntoufa et al., 2012; Decker and Peschel, 2001; Decker et al., 2000).
In the original paper of Arvaniti et al. (2011), describing the data used in this study, the researchers explored the differences in the expression profiles of the TLR signaling pathway between different subgroups of CLL patients, including M-CLL vs U-CLL. In this type of study, the quantitative research is conducted on a certain number of patients, due to limitations such as patient availability and the final output is a variable representing gene expression. Though enough for meaningful biological conclusions, the number of the samples is low for statistical analysis. In a previous study the difference in the signaling pathway between M-CLL and U-CLL was explored with the use of structural equation models (SEM) (Tsanousa et al., 2013). Given the inadequate sample size, it was inevitable to model bigger pathways using SEM. Therefore the idea of exploring the correlation structures and their significance of differences between M-CLL and U-CLL was born. Comparison of correlation refers to tests suitable for comparing two correlation matrices as a whole (Jennrich, 1970) or tests that compare two or more correlation coefficients (Fisher, 1925; Zou, 2007) to detect significant differences. SEM is a more detailed way of exploring correlation or covariance structures and is widely used in social sciences, finance and medicine. SEM is suitable for analyzing multiple equations and for measuring latent variables. The definition of the initial model is quite important since it affects the following process of estimation and re-specification. There are no available methods for selecting variables in SEM procedures; therefore the initial model is usually constructed according to theoretical knowledge. In the current study, there are numerous path models that could be constructed that may exist as causal relations among genes, but could not be statistically confirmed with the present data. Thus, a heuristic variable selection method based on the results of comparison of correlations is introduced, to proceed the construction of a model in structural equation modeling. The comparison of correlation coefficients provides a novel way to select which variables to include in the multigroup analysis with SEM, in order to further explore the differences of the specific gene associations. Therefore instead of studying the covariance structure of all variables included, we model the variables participating in correlations that significantly differ between the two patient groups. The comparison of correlation coefficients, as well as the study of correlation structures alone, could serve as a basis for developing new methodologies for feature selection or feature extraction. In summary, the contribution of the current work lies in the following: a) we propose utilizing the results of comparison of correlation coefficients as a feature selection method, b) we combine Structural Equation Modeling with a feature selection method, namely the proposed one, c) we emphasize the usefulness of structural equation modeling for studies of biological
networks d) we utilize graph modeling to confirm the differential structure of the genes between two subgroups of CLL patients.

The paper is organized as follows: in Section 2 research works relevant to the biological or statistical parts of the paper are briefly reviewed while Section 3 describes the statistical methods used in the paper. In Section 4, description of the data and analysis of the results are provided. Section 5 concludes the paper with the discussion of the findings. Finally the conclusions are presented in Section 6.

2. Related work
To our knowledge, comparison of correlations hasn’t been applied to gene expression data in CLL yet. However, various statistical processes have been utilized in gene expression profiling studies concerning CLL, for example t-tests to compare the average expression of genes between M-CLL and U-CLL patients (Arvaniti et al., 2011) and unsupervised hierarchical clustering on various samples from individual patients to discover overriding effect (Herishanu et al., 2011). In the same paper principal components analysis was applied to normalized gene expression profiles. In (Fernandez-Martinez et al., 2017) a framework is presented which is quite similar to ours. From a set of microarray data of CLL samples, they identify the genes with the highest predictive power using Fisher's ratio and feature elimination and they study their correlation networks. In (Wiestner, 2003) there has been extensive statistical analysis applied on gene expression profiles. Firstly correlations of gene expression profiles between CLL and other cord blood B cells were calculated using Pearson coefficient and no relation was found. T-tests were used on logarithmically transformed expression ratios to distinguish between the two CLL mutation types, both on univariate and multivariate level. For the multivariate analysis, linear combination of samples was used. Finally, cross validation analysis was applied in order to identify markers for the subtypes’ classification. Unsupervised as well as supervised clustering are utilized in (Klein et al., 2001) to discover patterns in gene expression profiles.

In more general studies of gene expression data, not necessarily focused on the specific disease, the use of statistics and bioinformatics is wide. Clustering algorithms have been proven very valuable in answering important questions of medical researchers. In (Yeung et al., 2001) various clustering techniques, applied on gene expression data, are compared, with Gaussian mixture models found to be the most effective ones. Another overview of clustering algorithms for gene expression profiles derived from microarray data can be found in Kerr et al. (2008). A study which is more related to the current paper can be found in D’haeseleer et
al. (2000) where methods for causal connectivity between genes are under research, along with the investigation and comparison of clustering statistical algorithms. Larger and more complicated gene expression datasets allow for the application of more complex algorithms from the field of bioinformatics that reveal patterns in data and can also be used for prediction based on clustering. Examples of such algorithms that utilize correlations or similar statistical measures like mutual information are CLR (Faith et al., 2007), ARACNE (Margolin et al., 2006) and MRNET (Meyer, 2008). Correlations are also utilized in feature selection algorithms, known as correlation-based feature selection algorithms (CBFS), which have been widely applied in machine learning applications especially of time series data, such as EEG signals (Mursalin et al., 2017). In CBFS algorithms, the correlation of each feature and the categorical target variable is calculated and if the coefficient is above a cut-off value, the feature is considered to contribute to the classifier algorithm and thus it is included in the model. In Kamaraj & Purohit (2016), cross-correlation analysis was utilized to analyze the mutations in MATP gene, that is related to albinism type IV. Cross-correlation analysis is also found in Purohit (2014), where it is applied to analyze the collective motions of a receptor which is important in the treatment of gastrointestinal cancer. The above mentioned references, chosen from many relevant studies, are indicative of the utility of correlation related methods for analysis of structures.

In most of the above mentioned references, the purpose was to find groups in data. In this paper, the discovery of significant groups is not of interest, since the two distinctive patient groups are already known. The important aspect is to reveal correlation patterns so to better understand the nature of the structural diversity in CLL. As a general idea it could be considered similar to that found in Faith et al. (2007), Margolin et al. (2006) and Meyer et al. (2008), since it uses comparison of correlations, instead of mutual information, to discover genes that could be utilized as predictors for the type of the disease. Due to the different nature of the data, bioinformatics’ tools used in the aforementioned papers, such as ARACNE and MRNET algorithms, could not be applied here and receive the results our suggested methodology derives.

3. Methods
3.1 Correlation coefficients
A correlation coefficient determines the bidirectional relationship between two variables. It gives information about the statistical significance of the relationship, the magnitude of the relation and the direction of it. The most commonly used correlation coefficients are Pearson's
r (Pearson, 1895) and Spearman's rho (Spearman, 1904). When variables are normally distributed, Pearson's correlation coefficient is applied, while Spearman's rho is preferred when the assumption of normality does not hold. The formula for Pearson’s correlation coefficient, which is applied here, in a sample of size \( n \) is given in

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

Eq. 1

Where \( x_i, y_i \) are the values of the normally distributed variables \( X, Y \) respectively, and \( \bar{x}, \bar{y} \) are the corresponding averages (Pearson, 1895).

3.2 Graphical modeling

Graph modeling is quite often employed in biological sciences, where there is a great need to study biological networks. The introductory terms needed for comprehending graph theory are nodes, edges and arcs. Nodes or vertices are the basic objects of a graph, usually represented by a circle. In this study nodes respond to genes. Edges are lines connecting unordered pairs of nodes, thus creating an undirected graph. Arcs on the other hand, connect ordered pairs of nodes, which constitute a directed graph (Biggs et al., 1976).

In the current study, in order to illustrate and emphasize the differences in the gene network structure of M-CLL subgroup versus the U-CLL subgroup, we used two types of undirected graphs that examine the pairwise relations between genes in terms of covariance. A model selection procedure precedes the visualization stage for each graph. In graphical modeling, model selection focuses on determining the support (i.e. the nonzero elements) of the precision matrix in order to produce a graph that represents potentially meaningful relations. The precision matrix is defined as the covariance matrix inverse (Dodge, 2006). To overcome issues of collinearity it is better to use the regularized estimate of the precision matrix and since there is interest in different groups in data, the goal is to estimate multiple regularized precision matrices. To detect and visualize the strongest differences in each group specific precision matrix, optimal penalty parameters are needed, calculated through leave-one-out cross-validation (LOOCV) procedure. In each run of the algorithm a sample is kept out (LOOCV) and a negative log-likelihood score is calculated for each value of the penalty parameter. The value of the penalty parameter that gives the lowest score is considered optimal. To determine the support of the estimated precision matrix, the latter is transformed
into a partial correlation matrix. Support determination can be achieved in various ways. One
of them is thresholding on the absolute values of the matrix entries by defining a specific cut-
off value (de Leeuw et al., 2017). The two types of graphs utilized are: a) the differential
graph and b) the community plot.

The so-called differential graph visualizes a network with edges that are found exclusively for
each of the groups, in this case the two patients' subgroups (de Leeuw et al., 2017). As already
described, to define the unique edges, the algorithm is based on an estimation procedure of
the two groups' respective precision matrices over the same variables that results in sparse
partial correlation matrices that have retained values above a certain threshold. Therefore the
edges in the differential graph represent positive or negative partial correlations above a
certain threshold in terms of absolute value (Peeters et al., 2015).

Community plots visualize the community structure of a network. Communities are clusters
of nodes that are naturally more densely connected to each other than to other nodes in a
network. To detect these communities, the Girvan-Newman algorithm is applied (de Leeuw et
al., 2017). For all edges in the network, the algorithm calculates the edge betweenness, which
is defined as the number of shortest paths between pairs of nodes that run along an edge. By
removing the edges with high edge betweenness, the groups are separated from one another
and the underlying community structure of the graph is revealed. For more information we
refer the reader to Girvan & Newman (2002).

3.3 Comparing correlation matrices

Comparison of correlation matrices is the first step of examining the existence of differences
between correlation structures of two (or more) samples. The proper method of comparison is
selected based on how the samples are related, whether they are dependent and overlapping,
dependent and non-overlapping or independent. The procedure compares the entire matrices,
not testing individually differences of specific elements. Few techniques are available for
comparing correlation matrices. In the comparison of two independent samples, which is our
case, the most appropriate technique is the chi-square test proposed by Jennrich (1970).

Jennrich proposed an asymptotic chi square test to examine the equality of two correlation
matrices. The test statistic used in Jennrich has the “form of a standard normal theory statistic
for testing the equality of two covariance matrices with a correction term added” (Jennrich,
1970). The hypothesis tested was that two normal populations have the same correlation
matrices, without having equal standard deviations or means. At that time, the only available
test for equality of correlation matrices was Kullback’s test (Kullback, 1967) which according
to Jennrich, did not follow an asymptotic chi square distribution under the null hypothesis. Kullback’s test was a generalization of Jennrich’s test, because it could compare more than two correlation matrices. The basic mathematical relations involved in Jennrich’s test are the following: for two sample correlation matrices $R_1$, $R_2$ of independent samples with $n_1$, $n_2$ sizes, the test statistic used in Jennrich’s approach is given by the following equation

$$X^2 = \frac{1}{2} tr(Z^2) - dg(Z)S^{-1} dg(Z)$$

Eq. 3

Where

$$Z = c^{1/2} R^{-1} (R_1 - R_2), R = (n_1 R_1 + n_2 R_2)/(n_1 + n_2), c = \frac{n_1 n_2}{n_1 + n_2}$$

$dg(Z)$: the diagonal of a square matrix $Z$ written as a column vector, $S = (\delta_{ij} + \bar{r}_{ij}^2)\delta_{ij}$ with $\delta_{ij}$ the Kronecker delta, which in linear algebra denotes the entries of an $n \times n$ identity matrix and $\bar{r}_{ij}$

For a more detailed description we refer the reader to the original paper (Jennrich, 1970). This theory is implemented in the R package \textit{psych} (Revelle, 2015).

### 3.4 Comparing correlation coefficients

In advanced studies, where the goal is to compare the correlation structure of groups of multidimensional data, it is reasonable to try to identify statistically significant differences between two or more correlation coefficients. For the comparison of two correlation coefficients, the choice of the proper method depends again on whether the samples which the coefficients are derived from are related or not. In case of independent samples, there are two ways to conduct the test: a) Fisher’s $z$ procedure (Fisher, 1925) and b) Zou’s confidence intervals (Zou, 2007). When the samples are related, they can be overlapping or non-overlapping. The relevant theory can be found in Diedenhofen and Musch (2015). In our study, only independent samples are involved so we consider the aforementioned tests.

#### 3.4.1 Fisher’s $z$ procedure

Fisher suggested the transformation of a correlation coefficient in order to resolve the following issues among others: a) Testing if an observed correlation differs significantly from a given theoretical value, b) Testing if two observed correlations are significantly different.
The method is similar to that used to test the significance of a correlation coefficient. The transformation $z$ is given by the following formula:

$$z = \frac{1}{2}[\ln (1 + r) - \ln (1 - r)] \tag{Eq. 4}$$

$z$ is approximately normally distributed, which makes the calculations for the requested tests simple. Some properties or characteristics of $z$ are:

- While $r$ moves from 0 to 1, $z$ increases from 0 to infinity.
- When $r$ is small, $z$ is nearly equal to $r$, but as $r$ tends to 1, $z$ increases unboundedly.
- When $r < 0$, so does $z$.

In Fisher's example of comparing two correlations, besides the difference of the corresponding $z$-values, the standard error of $z$ is also crucial for the calculations. When the difference of $z$-values does not exceed twice the standard error, there is no evidence of significantly different correlations. Fisher's $z$ follows normal distribution approximately, even when the data are not normally distributed. The test statistic $Z$ for the difference is calculated by Revelle (2015).

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{Eq. 5}$$

### 3.4.2 Zou's confidence intervals

In hypothesis testing, decisions are based either on p-values or on confidence intervals. According to Zou (2007), use of confidence intervals should be preferred, since they provide the analyst with the ability to estimate the amount of the effect as well as the existence of it.

Support for this argument is also found in Olkin and Finn (1995), suggesting that confidence intervals are a more suitable method for comparison of correlations. Zou has proposed approaches for independent samples, overlapping and non overlapping correlations and for comparison of independent multiple squared correlations.

Before Zou's method, simple asymptotic (SA) methods, based on the central limit theorem, were available. For the case of two independent samples, the lower ($L$) and upper ($U$) limits for a $100(1 - a)$% confidence interval for the difference between correlation coefficients $r_1$, $r_2$ with respective sample sizes $n_1, n_2$, are given by
\[(L, U) = r_1 - r_2 \pm z_{a/2} \sqrt{\frac{(1-r_1^2)}{n_1} + \frac{(1-r_2^2)}{n_2}} \quad \text{Eq. 6}\]

Although SA confidence intervals are widely used, they require some conditions to be fulfilled: a) changes of the underlying parameter do not affect the sampling distribution of the estimated parameter and b) the sampling distribution is approximately standard normal. When these conditions do not hold, results may be quite inaccurate. Bootstrap confidence intervals were created as an improvement of the SA ones, although they still have some disadvantages (Zou, 2007). Zou introduced alternatives to SA methods, the modified asymptotic methods (MA), which create confidence intervals not so demanding in terms of computational time and difficulty as bootstrap confidence intervals. Furthermore, they provide more accurate results than SA confidence intervals in some cases. The performance of the proposed methods has been evaluated by simulation studies.

For two independent samples, which is the case of interest in the current study, the procedure is as follows: a) first create confidence intervals for each correlation coefficient, using the Fisher’s r to z transformation, by

\[l = \frac{\exp(2l') - 1}{\exp(2l') + 1}, \quad u = \frac{\exp(2u') - 1}{\exp(2u') + 1} \quad \text{Eq. 7}\]

Where \(l', u' = z \pm 1.96 \sqrt{\frac{1}{n-3}}\)

b) Then create the final confidence interval for the difference of correlation coefficients by:

\[L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2} \quad \text{Eq. 8}\]

\[U = r_1 - r_2 - \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2} \quad \text{Eq. 9}\]

Where \(r_1, r_2\) are the correlation coefficients, \((l_1, u_1)\) is the confidence interval for \(r_1\) and \((l_2, u_2)\) for \(r_2\). If zero is included in the final confidence interval, the null hypothesis is retained, meaning that there is no significant difference between correlations.
3.5 Structural equation models

Structural equation modeling (SEM) is a multivariate technique suitable for confirmatory and exploratory analysis. It has the basic advantages of examining simultaneously multiple relations and analyzing latent variables (Hair et al., 1998). Latent variables, also called factors, are theoretical concepts—most often found in social sciences—that cannot be observed directly, thus they are defined through other measurable variables (Byrne, 2010). Factor analysis is also a method for assessing latent constructs, but differs from SEM in the issue of simultaneous analysis of multiple equations. Due to the latter property, it is obvious that a variable can be exogenous (independent) in one equation and endogenous (dependent) in another equation. This property makes SEM more suitable for the study of correlation or covariance structures than for the creation of prediction models. All models can be algebraically represented through equations or graphically depicted through path diagrams (Byrne, 2010).

The analysis is conducted repetitively. It begins with the definition of the initial model, which should be identified for the analysis to continue, meaning that the number of variances and covariances should be equal to the number of estimable parameters (Byrne, 2010). Next, the model is estimated. The most common method is maximum likelihood (ML). ML bears the assumption of normally distributed data and that sets certain limitations. In case the data do not follow normal distribution, other methods are available, such as the generalized least squares estimators (GLS) and methods assuming elliptical distribution and asymptotically distribution free estimation (ADF) that has no distribution assumptions at all (Bollen and Long, 1993). Another important aspect of SEM is the discovery and suggestion of relations in the data. This is done with the use of the modification indices (MIs). Larger values of the MIs are translated as bigger improvement in the model fit (Byrne, 2010). Each relation added should be considered not only from a mathematical but also from a theoretical or intuitive perspective, as it may have no logical meaning. The removal of causal relations with insignificant estimates and the addition of others proposed by the MIs, guide the analyst towards respecification. The whole procedure described can also be used to test the same construct against multiple groups. The analysis will provide different significant estimates and modification indices, which after being utilized in the respecification process, there will be different final models for each group.

The fit of each model created in each step, is accessed by various indices. A satisfactory fit can be a reason to terminate the process of redefining the equations constituting the structural equation model. There are numerous fit indices available while the most well known are:
• **Goodness-of-Fit Index (GFI)** and **Adjusted Goodness-of-Fit Index (AGFI)**, belong to the class of absolute measures of fit. Their values range between 0 and 1, with greater values suggesting better fit. GFI compares the model covariance structure to the sample data covariance structure (Cheung and Rensvold, 2002) while AGFI takes also into consideration the degrees of freedom (Byrne, 2010).

• **Normed Fit Index (NFI)** is a relative measure of fit, calculating the discrepancy between the independence model- a model with no correlations between variables-and the proposed model (Hair et al., 1998). Although NFI is widely used, it has been reported to be sensitive for small sized samples and underestimating fit (Bearden et al., 1982).

• **Comparative Fit Index (CFI)** is an adjustment of NFI that takes into consideration the sample size proposed by Bentler in 1990. It also belongs to the family of relative measures of fit (Byrne, 2010).

• **Root Mean Square Error of Approximation (RMSEA)** reflects how well the proposed model would fit the population. To receive good fit, its value should be less than 0.05 (Hair et al., 1998).

Various statistical packages are available for SEM and latent variables’ analysis, with LISREL (Jöreskog and Sörbom, 1989) being the oldest one. SPSS AMOS (Arbuckle and Wothke, 1999) has the ability to draw path diagrams to depict the equations instead of writing them separately. Freely available statistical language R has many packages offered for SEM, such as SEM package (Fox, 2006), OpenMx (Boker et al., 2011) and Lavaan (Rosseel, 2012). There are also features for SEM analysis in Psych package (Revelle, 2015).

4. Results

4.1 Data

The two datasets included expression profiles of 74 genes participating in the TLR pathway-10 genes were missing from this array since they were not expressed in CLL cells. The M-CLL subgroup (n=124) was almost twice as big as the U-CLL subgroup (n=67). There were missing expression profiles for some genes in 24 patients. The relative gene expression was found using real-time quantitative PCR (Polymerase chain reaction) on commercially available PCR arrays (RT² Profiler PCR arrays from Qiagen) and the 2(-Delta Delta C(T)) Method (Livak and Schmittgen, 2001) for the quantification of gene expressions. Most
patients were male. Particularly in the U-CLL subgroup, male patients are almost twice as many as the female patients (Table 1). The data were collected by the Hematology department of Papanikolaou hospital in Thessaloniki. A more detailed description of the data can be found in the original paper (Arvaniti et al., 2011).

<table>
<thead>
<tr>
<th></th>
<th>M-CLL subgroup</th>
<th>U-CLL subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>68</td>
<td>42</td>
</tr>
<tr>
<td>Female</td>
<td>53</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 1: Frequency of gender in each subgroup

4.2 Application
The analysis was conducted with the statistical programming language R. Various packages, that are mentioned in the following paragraphs, were utilized for the application. The steps of the analysis are graphically presented in Fig.1 and are explicitly described below:

a) Two correlation matrices were calculated based on Pearson’s correlation coefficient, each for a separate group of patients (M-CLL and U-CLL). The bivariate correlations in each one of them were graphically represented for visual inspection.

b) The two correlation matrices were compared as a whole, using the Jennrich's test.

c) Since statistically significant difference was found in step (b), the analysis proceeded in the comparison of the individual bivariate correlation coefficients between groups. Fisher's procedure and Zou's confidence intervals were applied to compare the correlation coefficients of the two patient groups.

d) Several pairwise correlations were found to differ between M-CLL and U-CLL subgroup. The variables (genes expressions) that participated in most of the statistically different correlation coefficients were identified as most "correlation-discriminative", essentially most important for the subsequent modeling. The number of statistical differences that make a variable important was set arbitrarily to 20.

e) The network of the most important genes identified in (d) and their already known theoretical connections was represented by structural equation models, estimated and assessed for fitting using the M-CLL and the U-CLL datasets separately.

f) The output of the above procedure contains comparison results and two different SEMs, one for each group of patients. The results show the extent of the structural difference of the two groups.
Figure 1: Flowchart representing the analysis' steps.

4.2.1 Study of correlation structures
The first step was to create separate correlation matrices [74x74] for the two independent samples (M-CLL and U-CLL). The elements of each correlation matrix are correlation coefficients of pairs of genes. In order to calculate the correlations, Pearson’s correlation coefficient was used. Correlation matrices were created using the R package “psych” (Revelle, 2015), which has the additional ability to provide p-values for each correlation that can later be used in the comparison.
Figure 2: Correlation matrices for each group of patients, where $m_{ij}$ and $u_{ij}$ ($i,j = 1,\ldots,74$) are Pearson correlation coefficients of M-CLL and U-CLL subgroups respectively.

Before proceeding with the comparison of correlations, it would be interesting to look at some results of this step, which are visually represented in Fig. 3. The plots were created using the package “corrplot” in R (Wei, 2013). It can be observed that in the U-CLL data subgroup there are more strong correlations than in the M-CLL subgroup. In M-CLL we have a few more negative correlations (depicted by white circles), although the number is small in both subgroups. This is a first indication of different correlation structures between the two groups of patients that will be further examined below.
Figure 3: Visualization plot of correlation coefficients of the two groups of patients. Empty cells are for insignificant correlations. The black color depicts positive coefficients, while the white color is for negative coefficients. Bigger dots stand for stronger correlations. Genes depicted in each row/column can be found in Table A1 (Appendix A).

Another way of detecting the different relationships formed between genes in the two groups of patients is to visualize their covariance structure by differential graphs (Peeters et al., 2015). This graph shows only the unique relations for each group. The green edges represent the partial correlation elements found only in the U-CLL subgroup and the red edges the partial correlation elements found only in M-CLL subgroup. The cut-off value for thresholding on the absolute values of the matrix entries was set to 0.05. The cross validated negative log-likelihood score is based on the correlation matrix. As it can be seen in Fig. 4, there are much more relations among the genes in the U-CLL case than in the M-CLL case, a result also established from the rest of the analysis and is in concordance with the relevant literature.

Figure 4: Differential graph. Relations existing exclusively in the M-CLL subgroup are depicted with red lines and green lines depict the unique relations found in U-CLL subgroup. Solid lines represent positive partial correlations, while dashed lines represent negative ones.

The relations between genes separately for each group of patients can be visualized in Fig. 5 along with the communities created for each group of patients.
The same pattern regarding the density of correlations that was observed previously is also apparent with these graphs. For the M-CLL subgroup there are some genes left aside, not forming any community, while the communities that are formed among the other genes are clearer. On the contrary, in the U-CLL subgroup there is a high density of correlations, the communities created are overlapping and there are only three genes not participating in any group. So the difference in the correlation structure among genes in the two data subgroups is once again confirmed visually. Besides suggestions for further investigation of the communities created, already reported biological interactions can be found among genes of the same community. Just to mention a few, in the M-CLL case, in the community formed by TLR1, TNF, TLR9 and NFKBIA, TLR9 stimulation is known to cause production of TNF (Rozkova et al., 2010). In another community, HSPA1A, HSPD1, and BTK all belong in the signaling complex (Arvaniti et al., 2011). In the U-CLL case specific relations are hard to study since most communities seem to overlap and there are numerous connections among genes. Looking at the distinctive community of TICAM2, IRAK1, TIRAP and SARM1, they all belong in the signaling complex (Arvaniti et al., 2011). From the small community of ELK1, FADD, CASP8 and BTK there is biological interaction between FADD and CASP8 (Micheau et al., 2003).
4.2.2 Comparison of correlation matrices
Initially we compared the two correlation matrices to find if there is statistically significant difference, using the Jennrich’s test from the package “Psych” (Revelle, 2015). The results were a chi square value of 8.2163,3 with a p<0.001, confirming significant difference between the two matrices. Since the comparison of the correlation matrices revealed that they are not equal, we proceed with the comparison of the matrices’ elements in order to detect where the differences lie in particular.

4.2.3 Comparison of elements of the correlation matrices
The following step was to detect which of these correlations were significantly different between the two samples. Using the package “cocor” (Diedenhofen and Musch, 2015), we tested the differences for each pair of correlations of the two samples. For each test, Fisher’s z test and the Zou’s confidence intervals were calculated. Although Zou's intervals are supposed to be a more accurate method, in this case the results of both procedures coincide, however Zou provides the additional information of the magnitude of difference. By examining the results, we distinguished the genes that participated in 20 or more pairwise correlations that were significantly different: These genes were CHUK, HMGB1, IFNA1, IL6, IRAK2, IRF3, CD180, MAP3K7, with HMGB1 and CD180 being the more "important" ones. HMGB1 expression levels have been found to differ between CLL patients and control group (Jia et al., 2014), which probably justifies the appearance of this gene in most pairs that significantly differ between M-CLL and U-CLL patients. Contemporary study reveals the importance of CD180 as a biomarker in hematologic tumors (Yeh et al., 2017), while different expression of CD180 has also been found between CLL patients and healthy volunteers (Antosz et al., 2009).

Focusing on the above mentioned genes that participate in the most differences of correlation coefficients and by visual inspection of their status in Fig. 5, we see that most of the aforementioned genes have more interactions in the U-CLL case. To briefly mention some of the differences observed, HMGB1 belongs to a crowded community in both M-CLL and U-CLL cases, however its interactions differ between the two patient groups. CHUK is found in different gene interactions across the two groups. IFNA1 belongs to communities of different density in each subgroup. IL6 participates in more interactions in the U-CLL case.

4.2.4 Structural equation models
The genes participating in most differences were assumed to play an important role in discriminating mutational and unmutational patients. Therefore they were used to build a structural equation model that was assessed using multigroup analysis in order to detect differences in gene relations. The implementation was made using the R package “lavaan” (Rosseel, 2012). The initial model was common for the two groups and is graphically depicted as a path diagram in Fig. 6. The model was based on the genes participating in most differences and their already known theoretical connections amongst each other and the other genes included in the current study. Genes receiving effect are those that arrows result in, while genes generating the effect-or signal- are the ones that arrows begin from. The theoretical knowledge utilized to create the relations among genes is described in Arvaniti et al. (2011). Maximum likelihood was the method used for estimation and any missing values were treated with listwise deletion. From the first estimation of the model the differences between M-CLL and U-CLL groups were obvious. Different regression coefficients were found insignificant for each group and different relations were proposed from the MIs to be added in order to improve fit. We resulted in two final models, one for each group, that confirm the existence of variability in structures. The final models for each group are shown as path diagrams in Fig. 7.

Figure 6: Initial model for multigroup analysis. Errors are not included to prevent complexity of the path diagram.
Figure 7: Final models for U-CLL group (left) and M-CLL group (right). The light grey arrows are for the initial relations. Black arrows are the relations of the final model while red arrows stand for the relations that are common in initial and final models.

As it can be observed, most of the initial relations do not appear in the final models. Comparing the final models, in M-CLL subgroup IL6, NFkB1 and RELA genes do not present any significant interactions. Indicative references can assure that IL6 promotes the progression of the disease (Buggins et al., 2008).

The fit of each model is confirmed by the Chi-square test but mostly the fit indices. The null hypothesis tested is that the proposed model holds in the population. However, in structural equation modeling, the fit indices are of more importance, since the acceptance or rejection of the null hypothesis according to the p-value can be misleading (Byrne, 2010). More specifically, the final model for the M-CLL case has a chi-square value of 9.615 (p-value=0.725). The NFI equals 0.956 and CFI, its adjustment, is very close to 1. The absolute measures of fit, GFI and AGFI, are also indicating very good fit, with values of 0.979 and 0.941 respectively. RMSEA is 0, verifying the very good fit. For the U-CLL case the final model gives a chi-square value of 32.675 (p-value=0.05), while most of the indices indicate very good fit of the proposed model to the data (CFI=0.953, NFI=0.885, GFI=0.910). AGFI equals 0.765, which was expected due to the sample size of the U-CLL subgroup and RMSEA was found equal to 0.097, which though is not smaller than 0.05, is not enough to prove inadequate fit.
5. Discussion

Various methods suitable for pattern recognition have been applied to gene studies, giving insight to biological quests. In our opinion, correlation patterns and causal connectivity have not been given the proper attention in gene expression literature. With the current study we aim to give insight to CLL by applying comparison of correlations and SEM in order to identify patterns. This statistical approach is alternative to the bioinformatics algorithms that are applied in complex datasets.

The application was made on expressions of 74 genes of CLL patients. As mentioned earlier, the patients are categorized in two subgroups, the M-CLL group and the U-CLL group and the interest lies in defining which genes have different status between the two groups. There is already a published research study which describes how the data were produced and includes basic statistical analysis. The novelty of this work is that it examines the structure of data in order to detect genes with differentiating behavior between the groups. Identifying the genes that have different status between the two patient groups could assist in better comprehension of the disease, which could lead to developing targeted treatments.

The study of correlation and covariance structures, showed that there is a general differentiating behaviour of genes between the two groups of patients. More specifically, by focusing on the graphs, it is obvious that in M-CLL subgroup there are much less unique relations than in the U-CLL case and the communities created are less overlapping. U-CLL group on the contrary has a much more complicated structure, with many relations between genes and overlapping communities. Also relations are observed between the genes of different communities. We could safely infer that the connectedness of U-CLL is much higher than M-CLL.

Furthermore, some of the genes found in many significantly different correlation coefficients, are already known biomarkers for CLL and others' expression differentiates between M-CLL and U-CLL patients. The communities created include genes that belong to the same family or genes that are already reported to share interactions.

Structural equation modeling is a widely applied multivariate technique that has the benefit of examining simultaneously many relations. We are suggesting a combination of SEM with tests for comparison of correlations, in order to optimize the results that can be derived when studying correlation or covariance structures. Our application methodology initially detects the differences in correlations between two subgroups of CLL patients. The findings suggest
which pairs of variables contribute to the discrimination of the patients in these subgroups. Multigroup analysis with structural equation models is then applied to expand the findings of variations in structures from pairs of variables to multiple equations. Creating structural equation models by using the important findings of comparison of correlations, produced relations most of which are known to exist, while others are not found in relevant literature but still describe meaningful biological interactions. The modeling of already known signal maps may not be accurate in case of small sample sizes due to the inability to receive satisfactory fit. The purpose of this work was not to model the exact relations mentioned in previous works of the research group, but to give new insight by applying statistical methods not utilized on this type of data.

6. Conclusion
The statistical comparison of correlation coefficients has been able to identify variables that have different behaviors between the two patient groups, a result confirmed by structural equation models and graph theory. The variables that participated in many significantly different correlation coefficients were considered to have discriminative power between the two groups of patients. Based on these variables, a model was created and its structure was examined with multigroup analysis of SEM and with graph models. Both methods revealed different covariance structure among the variables between the two groups. The comparison of correlations could be combined with structural equation modeling, since they both deal with interrelated concepts; correlation and covariance respectively. The initial model in SEM is usually constructed based on theoretical knowledge. This, along with the method’s limitations regarding the number of cases according to the number of variables, results in rare applications of SEM in big data. Nevertheless the method could also be used in future studies in subgroups of genes identified by RNA sequencing technologies considered as biological meaningful i.e. genes of interest revealed after gene ontology or gene enrichment analysis that participate in a certain process/pathway. Combining SEM with a heuristic variable selection method such as the suggested, could extend their application to larger data by eliminating the need to study theoretically all the attributes included in a large dataset. The comparison of correlation coefficients could generally serve as a feature selection method. The combination of these methods could be expanded to other datasets not only from the field of biology, but wherever correlation patterns are of great importance. Suggestion for future work includes investigation of results of comparison of correlations in order to be used as a
feature selection method and combining comparison of correlations with other methods, especially in the area of data mining. Furthermore, an interesting direction for future research in CLL with the proposed methodologies, is the inclusion of other than the type of the disease factors that discriminate the patients, such as the gender.
References


Byrne, B. M. (2010). "Structural equation modeling with AMOS: Basic concepts, applications, and programming".


signaling, NF-κB activation, and tumor proliferation in chronic lymphocytic leukemia”. Blood, 117(2), 563-574.


Pathway in Chronic Lymphocytic Leukemia". In Database and Expert Systems Applications (DEXA), 2013 24th International Workshop on (pp. 71-75). IEEE.


Appendix A

Table A1: Names of genes in the order they appear in Fig. 3

<table>
<thead>
<tr>
<th>Number</th>
<th>Gene name</th>
<th>Number</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BTK</td>
<td>38</td>
<td>MAP4K4</td>
</tr>
<tr>
<td>2</td>
<td>CASP8</td>
<td>39</td>
<td>MAP8</td>
</tr>
<tr>
<td>3</td>
<td>CD80</td>
<td>40</td>
<td>MAP8IP3</td>
</tr>
<tr>
<td>4</td>
<td>CD86</td>
<td>41</td>
<td>MYD88</td>
</tr>
<tr>
<td>5</td>
<td>CHUK</td>
<td>42</td>
<td>NFKB1</td>
</tr>
<tr>
<td>6</td>
<td>EIF2AK2</td>
<td>43</td>
<td>NFKB2</td>
</tr>
<tr>
<td>7</td>
<td>ELK1</td>
<td>44</td>
<td>NFKB1A</td>
</tr>
<tr>
<td>8</td>
<td>FADD</td>
<td>45</td>
<td>NFKBIL1</td>
</tr>
<tr>
<td>9</td>
<td>FOS</td>
<td>46</td>
<td>NFRKB</td>
</tr>
<tr>
<td>10</td>
<td>HMGB1</td>
<td>47</td>
<td>NR2C2</td>
</tr>
<tr>
<td>11</td>
<td>HRAS</td>
<td>48</td>
<td>PELI1</td>
</tr>
<tr>
<td>12</td>
<td>HSPA1A</td>
<td>49</td>
<td>PPARA</td>
</tr>
<tr>
<td>13</td>
<td>HSPD1</td>
<td>50</td>
<td>PRKRL</td>
</tr>
<tr>
<td>14</td>
<td>IFNA1</td>
<td>51</td>
<td>REL</td>
</tr>
<tr>
<td>15</td>
<td>IFNB1</td>
<td>52</td>
<td>RELA</td>
</tr>
<tr>
<td>16</td>
<td>IFNG</td>
<td>53</td>
<td>RIPK2</td>
</tr>
<tr>
<td>17</td>
<td>IKBKB</td>
<td>54</td>
<td>SARM1</td>
</tr>
<tr>
<td>18</td>
<td>IL10</td>
<td>55</td>
<td>SIGIRR</td>
</tr>
<tr>
<td>19</td>
<td>IL12A</td>
<td>56</td>
<td>ECSIT</td>
</tr>
<tr>
<td>20</td>
<td>IL1A</td>
<td>57</td>
<td>TBK1</td>
</tr>
<tr>
<td>21</td>
<td>IL1B</td>
<td>58</td>
<td>TICAM2</td>
</tr>
<tr>
<td>22</td>
<td>IL6</td>
<td>59</td>
<td>TIRAP</td>
</tr>
<tr>
<td>23</td>
<td>IL8</td>
<td>60</td>
<td>TLR1</td>
</tr>
<tr>
<td>24</td>
<td>IRAK1</td>
<td>61</td>
<td>TLR10</td>
</tr>
<tr>
<td>25</td>
<td>IRAK2</td>
<td>62</td>
<td>TLR2</td>
</tr>
<tr>
<td>26</td>
<td>IRF1</td>
<td>63</td>
<td>TLR4</td>
</tr>
<tr>
<td>27</td>
<td>IRF3</td>
<td>64</td>
<td>TLR6</td>
</tr>
<tr>
<td>28</td>
<td>JUN</td>
<td>65</td>
<td>TLR7</td>
</tr>
<tr>
<td>29</td>
<td>LTA</td>
<td>66</td>
<td>TLR8</td>
</tr>
<tr>
<td>30</td>
<td>CD180</td>
<td>67</td>
<td>TLR9</td>
</tr>
<tr>
<td>31</td>
<td>LY86</td>
<td>68</td>
<td>TNF</td>
</tr>
<tr>
<td>32</td>
<td>LY96</td>
<td>69</td>
<td>TNFRSF1A</td>
</tr>
<tr>
<td>33</td>
<td>MAP2K3</td>
<td>70</td>
<td>TOLLIP</td>
</tr>
<tr>
<td>34</td>
<td>MAP2K4</td>
<td>71</td>
<td>TRAF6</td>
</tr>
<tr>
<td>35</td>
<td>MAP3K1</td>
<td>72</td>
<td>TICAM1</td>
</tr>
<tr>
<td>36</td>
<td>MAP3K7</td>
<td>73</td>
<td>UBE2N</td>
</tr>
<tr>
<td>37</td>
<td>MAP3K7IP1</td>
<td>74</td>
<td>UBE2V1</td>
</tr>
</tbody>
</table>
Highlights

- We utilized comparison of correlations of genes between M-CLL and U-CLL patients
- We applied SEM and graph modeling to study the differentiating structures
- Stronger correlations are found in U-CLL patients
- Community plots reveal more dense and interconnected communities in U-CLL cases
- HMGB1 and CD180 are in most differences of correlations between U-CLL and M-CLL
Graphical abstract