EXTRACTION OF DYNAMIC PATTERNS FROM STATIC RNA EXPRESSION DATA: AN APPLICATION TO HEMATOLOGICAL NEOPLASMS

Relatore: Barbara Di Camillo

Correlatore: Dott.ssa Alessandra Trojani

Laureando: Giulia Bianchi

ANNO ACCADEMICO 2012/2013

Nothing worth gaining
was ever gained without effort.

Theodore Roosevelt

Index

Section 1

INTRODUCTION

Development and evolution of a disease are dynamic processes that, from a molecular point of view, involve changes in some gene expression levels in the involved organs and cells. A possible approach to study the behavior of such dynamic phenomena is to sample individuals, tissues or other relevant units at subsequent time-points throughout the progression. In this thesis we focused on two pathological conditions: chronic lymphocytic leukemia (CLL) and Waldenström's macroglobulinemia (WM). In the first case we sought genes responsible for different prognosis of CLL and tried to classify patients with an undefined prognosis. In the latter case we sought genes responsible for the evolution of IgM monoclonal gammopathy of undetermined significance (IgM MGUS) in Waldenström's macroglobulinemia.

To gain our goals, we used a tool, Sample Progression Discovery (SPD), developed by Peng Qiu et al. (1). This software, given gene expression data, extracts those features that, by gradually changing their expression values throughout samples, are responsible for leading some biologically meaningful process. A progression is not necessarily temporal: can also represent the disease evolution, or any other kind of progression, provided an ordering criterion was previously specified.

CLL has been widely studied and two prognosis classes have been defined, based on two biomarkers: the mutational status of IgV_H and the expression of ZAP70. An un-mutated status of IgV_H together with positive expression of ZAP70 is correlated to a poor prognosis and the need of treatment. On the other hand, a mutated status of IgV_H and negative expression of ZAP70 is related to ~~a~~ positive prognosis and no treatment is provided. Some uncertainty arises when, in a patient, the two biomarker values are such that the classification into one of the prognostic classes is not possible, i.e. the patient shows mutated IgV_H and ZAP70 positive expression. Hence, we tried to establish how to consider such undefined cases.

Regarding MGUS and WM, recent studies showed that MGUS is the most common plasma cell dyscrasia and is associated with a lifelong risk of progression to multiple myeloma or related disorders (2). Thus we used SPD to have a better understanding on such evolution and the involved genes.

After a first part of the thesis, during which we had to evaluate how to set SPD parameters and how to preprocess data, we went into a second part. Indeed, high-throughput expression data are affected by noise as the number of genes is great ($\sim 5 \cdot 10^4$), whereas the number of samples for each patient is one or two. In the latter part, we actually applied the method to reach our aims.

1.1. Extraction of temporal dynamics from gene expression data

High-throughput expression data can be used to infer temporal orderings by assuming that the process evolution can be detected by relatively smooth and continuous changes in the transcriptome (3).

The estimate of accurate time series for biological processes is a hard task, due to the complexity of the problem. First of all, it is not always linear. For example, in the formation of blood cells, hematopoietic stem cells can differentiate in both myeloid and lymphoid cells. The latter two are the starting point of several parallel pathways, eventually leading to all blood cell types (4).

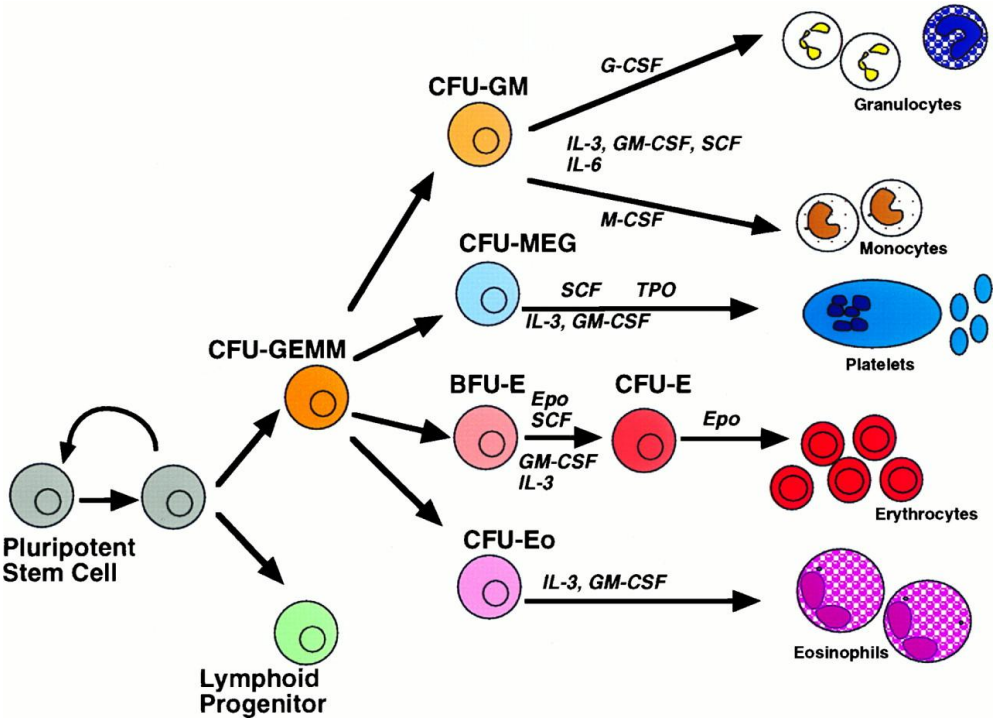

Figure 1

Besides the process itself, a further issue concerns the experiment to extract genetic material from cells for gene expression analysis. Indeed, time series data are usually drawn from a population of cells and, if they are not synchronized, samples can contain mixtures of the

temporal process. Moreover, heterogeneity among the members of the population further complicates the temporal samples and can lead to ambiguous situations, in which sample orderings are correct with respect to absolute time, but don't follow the dynamics of the biological process.

1.2. Chronic Lymphocytic Leukemia

Chronic Lymphocytic Leukemia (CLL) is a chronic inherited lymphoproliferative disorder. It is the most common type of leukemia in Western countries, and is characterized by an increasing amount of mature-looking immuno-incompetent lymphocytes; 95% being B cells. The amassment of the clonal cell population occurs in blood, bone marrow, lymph nodes and spleen. The diagnosis of CLL is conventionally set in the presence of more than 5,000 small mature-appearing lymphocytes per μl of peripheral blood. It is more common in males than in females, and affects especially elderly people. The etiology of CLL is still being investigated, as it is not known yet, but seems reasonable that the genetic predisposition may be the best explanation. In fact there is a higher prevalence of the disease in the family of the patients and there is no established role of the environment as inducing or influencing factor (5).

As a matter of fact, CLL is a disease with a highly variable course, mainly falling into two subclasses, the most important difference being the illness aggressiveness. For patients in whom the disease has a slow course, there is no need for a specific therapy, whereas, for those suffering from a more aggressive pathology, treatment is urgently needed. CLL patients presenting leukemic cells that have rearranged genes coding for the variable region of the heavy chain of sIg (IgV_H) with more than 2% mutations are in general considered good prognosis patients, whereas those ones who do not show IgV_H mutations have in general worse prognosis (6). As the DNA sequencing to determine the status of IgV_H mutation is expensive and not usually performed in all clinical contexts, several studies aimed to find alternative factors and biomarkers that can be correlated to such a difference.

Nowadays, one well established of such markers is ZAP70, an intracellular protein that triggers activation signals delivered to T lymphocytes and natural killer cells by surface receptors for antigens. It is rarely present in normal B cells, but has been found in B cells from patients with CLL (6).

Indeed, DNA analysis on gene expression of B cells showed ZAP70 expression to be strongly associated with mutational status of the IgV_H gene. More specifically, further analysis confirmed that ZAP70 positivity is related to an un-mutated IgV_H gene status, whereas ZAP70 negativity is

associated with a mutated IgV_H gene status (7). Thus, according to the combination of these two conditions, patients can be stratified into two prognostic groups (8):

- positive prognosis, characterized by mutated IgV_H and ZAP70 negative (M-ZAP70$^-$);
- poor prognosis, characterized by un-mutated IgV_H and ZAP70 positive (UM-ZAP70$^+$).

Other independent molecular markers in CLL include surface CD38 expression and the presence of specific chromosomal aberrations, such as trisomy 12 and 11q22-23, 13q14, 6q21, 17p13 deletions (5). More recently, lipoprotein lipase (LPL) expression has been shown to correlate with IgV_H mutational status (9), as well as deregulation of expression of genes coding for enzymes controlling lipid metabolism (8).

Prognosis	Biomarkers	
	IgV_H mutational status	ZAP70 expression
POSITIVE	mutated	-
POOR	un-mutated	+

Table 1

1.3. IgM Monoclonal Gammopathy of Undetermined Significance and Waldenström's Macroglobulinemia

MGUS is an asymptomatic premalignant disorder characterized by limited monoclonal plasma cell proliferation in the bone marrow and absence of end-organ damage. MGUS is differentiated from multiple myeloma and related disorders based on the presence or absence of end-organ damage that can be attributed to the plasma cell disorder. It is characterized by a serum IgM concentration lower than 3.0 g/dL, infiltration of clonal plasma cells in the bone marrow lower than 10% and the absence of end-organ damage. Among all MGUS cases, approximately 15% involves serum IgM paraprotein. It is more common in men than women and in whites than blacks. Patients diagnosed IgM MGUS have an increased risk of Waldenström's macroglobulinemia and, therefore, IgM MGUS is considered a precursor of WM (10).

WM is a clonal IgM monoclonal protein-secreting lymphoid and plasma cell disorder. It is defined as an IgM monoclonal gammopathy with infiltration of clonal plasma cells in the bone marrow greater than 10%. Smoldering Waldenström's macroglobulinemia (also referred to as indolent or asymptomatic WM) is defined as serum IgM monoclonal protein level greater or equal to 3g/dL and/or bone marrow lymphoplasmacytic infiltration greater or equal to 10% and

no evidence of end-organ damage, such as anemia, constitutional symptoms, hyperviscosity, lymphadenopathy, or hepatosplenomegaly (2).

Patients with IgM MGUS and smoldering WM have an overall survival rate similar to the general population and should not be considered to have a malignant disease.

Diagnosis	IgM concentration [g/dL]	Bone marrow infiltration [%]	End-organ damage
IgM MGUS	<3	<10	no
WM	<3	<10	yes
sWM	≥3	≥10	no

Table 2

Section 2

DATA

As previously explained, this thesis is divided into two parts.
The main objective of the first part was to understand how to handle SPD. More specifically, we faced two main issues: selection of differentially expressed genes and the parameter setting. These two problems are closely connected, as they both affect the number of genes that are actually used by SPD to extract a progression.
To come up with an appropriate strategy of gene selection and suitable parameter values, we proceeded in the following way. Starting from microarray expression data belonging to patients suffering from CLL with known prognosis, we made five different selections of genes, as further explained later on, and tested SPD by checking if it could classify correctly patients in the two prognostic groups. For each of the five selections provided to SPD, we set the standard deviation threshold within a range of values and evaluated the classification error for each of them.

During the second part of the thesis, we used two other data sets to obtain different kind of information for two different diseases: CLL and WM.
Concerning CLL, we used the same data set used for the first part with a couple of differences: two microarrays have been added; a third class of patients with uncertain prognosis has been taken into account, as the objectives of this part were to evaluate how to treat undefined prognosis patients and try to classify one patient for which IgV_H mutational status was not available.
Regarding WM, we used a totally different data set. We analyzed microarray data belonging to 97 patients. Twenty-five of them were diagnosed IgM MGUS, the remaining 72 with WM. Samples were taken from different cell types: CD19 antigen positive, CD138 antigen positive and antigen negative.

2.1. CLL Data set

We examined microarray expression data belonging to patients affected by Chronic Lymphocytic Leukemia, diagnosed at the Division of Hematology at Niguarda Hospital, Milan, Italy. They were submitted to a software application called Sample Progression Discovery (1) to extract a progression underlying gene expression data. Microarray data belong to 112 patients divided into three classes: class 1 with mutated IgV_H and $ZAP70^-$, class 2 with un-mutated IgV_H and $ZAP70^+$, class 3 including both un-mutated IgV_H and $ZAP70^-$ and mutated IgV_H and $ZAP70^+$. Peripheral blood mononuclear cells (PBMCs) from all samples were isolated by Ficoll density gradient centrifugation (Invitrogen, Milan, Italy) at 800 rpm for 20 minutes and soon after $CD19^+$ cells were purified using MACS CD19 Microbeads (Miltenyi Biotech, Bologna, Italy) from fresh PBMCs of all 112 CLL patients following the manufacturer's instructions; the purity of $CD19^+$ cell was greater than 97% as determined by flow citometry. $CD19^+$ cells ($1 \cdot 10^7$) were resuspended in 100 µl of RNA*later* (Ambion, Applied Biosystems, Milan, Italy) and stored in a CLL cell bank at -20°C until RNA extraction was performed (8).

2.2. WM/MGUS Data set

This data set consisted of 97 probes belonging to patients diagnosed either with WM or with IgM MGUS. Bone marrow $CD19^+$, $CD138^+$ and NEG cells were isolated from WM patients and IgM MGUS patients as shown in Table 3. Bone marrow mononuclear cells from all samples were isolated by Ficoll density gradient centrifugation at 800 rpm for 20 minutes. Right after $CD19^+$ cells were selected using MACS CD19 Microbeads (Miltenyi Biotech, Bologna, Italy); afterwards $CD138^+$ cells were positively isolated from the collected $CD19^-$ cells using MACS CD138 Microbeads following the manufacturer's instructions (Miltenyi). Gene expression profiling has been performed on total RNA extracted from bone marrow $CD19^+$, bone marrow $CD138^+$ and NEG cells. The concentration and quality of the RNA samples have been evaluated using Nanodrop 2000 UV-vis spectrophotometer (Euroclone, Milan, Italy).

<table>
<tr><td rowspan="6">97 probes</td><td rowspan="2">38 $CD19^+$</td><td>27 WM</td><td rowspan="6">72 WM
25 IgM MGUS</td></tr>
<tr><td>11 IgM MGUS</td></tr>
<tr><td rowspan="2">31 $CD138^+$</td><td>24 WM</td></tr>
<tr><td>7 IgM MGUS</td></tr>
<tr><td rowspan="2">28 NEG</td><td>21 WM</td></tr>
<tr><td>7 IgM MGUS</td></tr>
</table>

Table 3

Section 3

SAMPLE PROGRESSION DISCOVERY

Peng Qiu et al. developed a tool, called Sample Progression Discovery (SPD), which aims to reveal biological progression underlying a microarray data set.

Based on the hypothesis that to any step of biological progression corresponds a gradual change in expression levels of some subsets of genes, SPD assumes that individual samples of a microarray data set are linked by an unknown biological process, and that each sample represents one unknown point along the progression of such process. So SPD can be useful when microarray samples are available but their ordering is unknown and not necessarily linear. In the latter case, SPD can detect branching points along the progression. It also has a feature selection ability that enables to reveal the candidate genes that regulate that progression.

SPD was tested on a variety of biological processes such as differentiation, development, cell cycle and disease progression. Microarray data sets, obtained by sampling a biological process at different points along its progression, were provided to SPD, without any other piece of information regarding either the progression itself or meaningful gene features.

3.1. Methods

SPD is implemented in Matlab 7 (The MathWorks, Natick, MA) using a graphical user interface. The software, at its second version, is available and freely downloadable on the internet at http://odin.mdacc.tmc.edu/~pqiu/software/SPD/.

The algorithm executes four main steps on gene expression data to finally extract a disease progression from expression data.

At first, co-expressed genes are grouped together into modules via clustering. This is necessary to speed up the analysis. The method used to cluster and obtain consistent gene modules is a consensus k-means agglomerative algorithm. The stopping criterion is the desired module coherence that is computed as the average Pearson correlation between each gene in the cluster and the cluster mean, and it is chosen by the user.

Information about the behavior of genes in the same module is translated into a minimum spanning tree (MST): each MST represents a group of highly co-expressed genes following the same pattern. Each node is a microarray sample and the edges are weighted by the distance between sample gene expression profiles. By definition, given a connected, undirected graph, a spanning tree is a sub graph that is a tree and connects all the vertices together. If a weight is assigned to each edge, the minimum spanning tree is the ST with weight less than or equal to the weight of every other ST. Hence, a MST connects samples that are closer to each other. Using such a structure to describe a progression enables SPD to find progressions with branching points, as well as linear ones.

Modules that share common MST structure are selected; the overall MST, representing the global progression, is pieced together using only the genes belonging to selected modules.

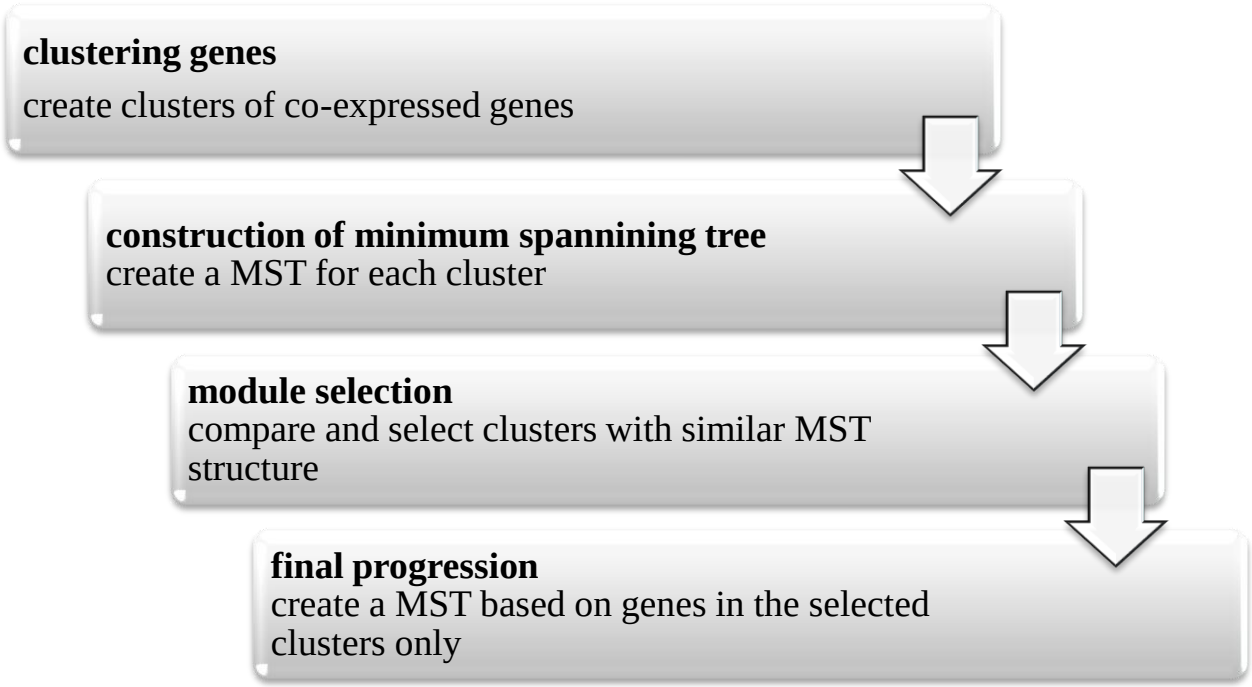

Figure 2

To assess the resemblance of progression patterns, SPD compares modules and MSTs. More specifically, it computes the earth mover's distance between all the modules and all the genes. This is one of the major differences between the latter version of SPD and the former one. To

identify similar modules in terms of progression, SPD generates a similarity matrix. It is necessary to choose a threshold that determines whether the fit between a module and a tree is significant. This parameter is user defined and its default value is 0.05: it means that, among all the module-tree pairs, the top 5% with most significant earth mover's distances are considered to "fit well with each other", and are used to construct the PSM.

The selection of modules needed to obtain the overall MST, thus the sought progression, is done manually. The user decision is supported by the progression similarity matrix. Each element represents the number of trees that are concordant with the two correspondent modules. Genes belonging to the selected modules are the ones supporting the overall progression. Hence, the feature selection is made by evaluating the statistical concordance between each gene module and each MST.

3.2. SPD step by step

3.2.1. Input format

To start using SPD is necessary to prepare a file *.mat* which contains at least three variables:

- probe_names: N·1 cell array with the N names of genes or features;
- exp_names: 1·M cell array with the M names of samples or arrays;
- data: N·M matrix of expression data.

Two optional variables can be added for the color coding of results:

- color_code_names: K·1 cell array with a name for each desired color code;
- color_code_vectors: K·M matrix with the clinical info corresponding to each color code.

Once the input file is ready, it can be loaded. It is also possible to load previous results.

3.2.2. Gene filtering

SPD gives the users three options to filter genes: standard deviation threshold, number of acceptable nulls per gene and throw away "_x_at" probes. If more appropriate filtering is needed, users can customize their own selection when preparing the input file, and then ignore the filtering options in the GUI.

The aim of the first option is to keep only those genes that are differentially expressed along the probes for the following analysis. In fact SPD computes for each gene the standard deviation σ on the expression values and all the genes with $\sigma<\sigma_{th}$ are not taken into account.

The second and the third options allow the user to "clean" the data set by removing all the genes with more NULL entries than acceptable and the "_x_at" probes. In the Affimetrix U133a GeneChip the "_at" suffix designates a unique probe set, while "_s_at" and "_x_at" suffixes designate probe sets that can cross hybridize with multiple genes (11).

The software shows the basic information after filtering.

3.2.3. Clustering

The aim of clustering is to group together highly co-expressed genes into modules so that the number of gene expression patterns to be tested is reduced. The algorithm chosen to cluster is an iterative consensus k-means procedure.

A k-means clustering algorithm performs the following steps:

1. randomly select k initial centers;
2. assign each element to the closest center according to a chosen metric;
3. re-calculate centers;
4. repeat 2 and 3 until a stopping condition is reached.

The number of clusters k has to be previously decided. SPD uses k=2 and iterates the algorithm for L=200 times.

Starting from the N·M expression data matrix, it creates a N·L matrix in which the (i, j) element represents the cluster assignment of gene i at iteration j. To have the consensus, k-means is applied again on the N·L matrix.

At this point, SPD creates modules by further clusterization based on the coherence of already existing clusters. The coherence is computed as the average Pearson correlation between each gene in the cluster and the cluster mean. If the coherence is less than a pre-specified threshold, such cluster is partitioned by iterating the procedure until all the clusters have coherence greater than the threshold that is set equal to 0.9.

The modules obtained so far, are not the ultimate ones. To be sure they are not similar to each other, they are compared pairwise. If the Pearson correlation of two modules centers is higher

than a user defined threshold, the modules are merged together. The suggested value for module coherence is 0.7 but, if the histogram of all the pair-wise correlations shows a heavy tail, a higher coherence parameter may be more appropriate.

3.2.4. Construct MSTs – Compare modules and MSTs

SPD builds a minimum spanning tree for each module, based on expression data of genes belonging to the same module.

Given a connected, undirected graph $G = (V, E)$, with V the set of vertices and E the set of edges, for each edge $(u, v) \in E$ a cost function $w: E \rightarrow \mathbb{R}$ is defined. A minimum spanning tree is an acyclic subset $T \subseteq E$ that connects all of the vertices and it is such that the total weight $w(T) = \sum_{(u,v)\in T} w(u, v)$ is minimum.

All the minimum spanning trees built on gene modules have samples as vertices and the weight of an edge connecting samples (u, v) is defined as the Euclidean distance between gene expression profiles of sample u and v. This way, MSTs connect samples that are similar to each other, but they are slightly different for a gradual change in gene expression.

The comparison between modules and trees is made by using the earth mover's distance (EMD) as a metric to build a progression similarity matrix. The goal is to seek for statistical concordance between all the modules and all the trees and, hence, determine the progression supported by meaningful features.
The earth mover's distance is the extension of the notion of distance between single objects to distance between distributions. Given two distributions, one can be seen as a mass of earth spread in the space, the other one as a collection of holes in the same space. If needed one can switch what is called earth and what is called holes so that there is always at least as much earth as needed to fill all the holes completely. The EMD measures the least amount of work necessary to fill the holes with earth, where work corresponds to transporting a unit of earth by a unit of ground distance (12).

The computation of EMD is based on the *transportation problem*. It deals with suppliers and consumers: suppliers have sources available to satisfy the consumer's demand. The goal is to minimize the cost for shipping the sources. This is a bipartite network flow problem, since the nodes can be divided into two parts with all arcs going from one part to the other (13), as shown in Figure 3, which represents an example with three supplier and two consumers.

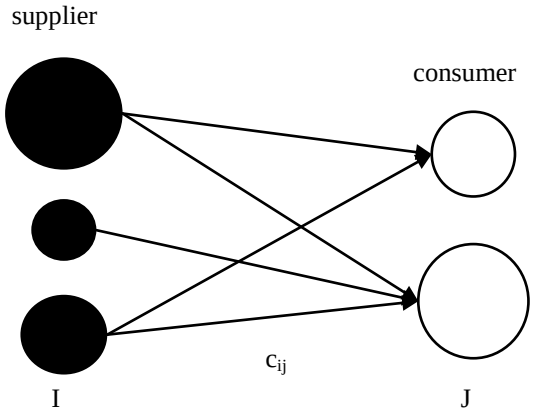

Figure 3

The problem can be formalized as the following linear programming problem: let I be a set of suppliers, J a set of consumers, and c_{ij} the cost to ship a unit of supply from $i \in I$ to $j \in J$. The solution is a set of flows f_{ij} that minimize the overall cost

$$\sum_{i \in I} \sum_{j \in J} c_{ij} \cdot f_{ij}$$

subject to the following constraints:

$$f_{ij} \geq 0$$

for $i \in I, j \in J$;

$$\sum_{i \in I} f_{ij} = y_j$$

for $j \in J$, where y_j is the total capacity of consumer j;

$$\sum_{j \in J} f_{ij} \leq x_i$$

for $i \in I$, where x_i is the total supply of supplier i.

The first constraint allows shipping of supplies from a supplier to a consumer and not vice versa. The second constraint forces the consumers to fill up all of their capacities and the last constraint

limits the supply that a supplier can send to its total amount. A feasibility condition is that the total demand does not exceed the total supply:

$$\sum_{j \in J} y_j \leq \sum_{i \in I} x_i$$

Once the transportation problem is solved, the earth mover's distance is defined as:

$$EMD(x, y) = \frac{\sum_{i \in I} \sum_{j \in J} c_{ij} \cdot f_{ij}}{\sum_{i \in I} \sum_{j \in J} f_{ij}} = \frac{\sum_{i \in I} \sum_{j \in J} c_{ij} \cdot f_{ij}}{\sum_{j \in J} y_j}$$

In general, the ground distance c_{ij} can be any distance and it will be chosen according to the problem to handle (12).

3.2.5. Identify modules similar in terms of progression

The main step of SPD that enables the extraction of the final progression is the comparison between the expression of gene modules and trees constructed from other modules. Based on the statistical concordance between al the modules and all the trees, which is user defined, a progression similarity matrix is derived. From this matrix, similar modules are easily recognizable, because they lie on the diagonal and they have red shade, darker or lighter depending on greater or smaller similarity. Since the number of modules in the progression similarity matrix is usually small, the module selection is to be performed manually.

Given the expression data of a gene module in M samples, a M·M distance matrix D is defined, where D_{ij} is the EMD distance between the gene expression profiles i and j. A second M·M matrix A is defined to represent the tree structure. It is an adjacency matrix in which A_{ij}=1 if samples i and j are directly connected in the tree, otherwise A_{ij}=0.
The concordance between a gene module and a tree is then defined as the concordance between the distance matrix D and the adjacency matrix A:

$$s = \sum_{A_{ij}=1} D_{ij}$$

In this way, the distance on the progression between connected samples is small whereas the distance between not connected samples is relatively greater. To derive the p-value of s, SPD performs 1000 random permutations. The threshold to compare the p-value is user defined, thought the suggested values to use are 0.05 (the default), 0.10 and 0.15.

Once the user chooses similar modules by visual inspection of the progression similarity matrix, SPD creates the overall progression based on genes belonging to those modules.

3.3. Results and discussion

SPD underwent trials to prove its potentiality of retrieving biological processes given microarray samples. The authors prepared several microarray data sets to test SPD. For each data set, the actual underlying progression was known, but was not provided to SPD, and was used merely as a comparison to validate the results.
The data sets included a cell cycle time series, B-cell differentiation data and a prostate cancer microarray data set. In all of the trials, SPD was able to recover respectively the correct time order of the samples, the correct order of different stages of normal B-cell differentiation, and the progression consistent with disease evolution. Moreover, the genes identified and involved to assess the progression were consistent with the biological process itself.

SPD has some distinctive features which make its analysis on microarray data more complete than other previously adopted techniques. Unlike other machine learning algorithms, such as unsupervised clustering, supervised classification and statistical tests for differential expression, whose goal is to identify discrepancies between different sample groups by assuming that samples in the same group are similar, SPD is based on an alternative approach. As it considers individual samples as different points along an unknown biological progression, it has the potential to discover how samples progress both within and across groups.
Furthermore, it is capable to extract the genes associated with the progression with no *a priori* knowledge about meaningful gene features.
One of the key aspects of this tool is the way used to measure the similarity among gene modules. Unlike methods using correlation and regression, in which the expression profiles of gene modules are directly compared with each other, in SPD the comparison is assessed via minimum spanning trees. MSTs represent progression patterns and the similarity between gene modules is based on the number of MSTs they share.
This way, SPD can identify similarities that correlation and regression-based analysis may miss.

Section 4

PARAMETER SETTING

To study the evolution of chronic lymphocytic leukemia, we decided to use SPD as a tool to assess the progression of the disease. First of all, we needed a method to decide how to set parameters in SPD. As here in below explained, results may greatly vary depending on such parameters; hence we tested SPD on 5 groups of genes. Each group was obtained via different methods of selection, by coding with R language (14).

4.1. Input configuration

We provided SPD with different sets of data. All were from the same microarray data set of a cohort of 112 patients, each group differing from the others for the gene selection. Out of the 112 patients, only those with a known prognosis were considered, namely 89.

The data set had to be normalized because samples were taken at different times, so we multiplied each column by a scale factor (columns represent patients, rows represent probes). We computed the median for each column and then computed the overall median (median of the medians). Each scale factor was given by the overall median divided by the median of each column. Once data were homogeneous, we ordered patients according to their class and deleted those belonging to class 3. We obtained a matrix with N=54675 probes on the rows and 89 patients on the columns, of which the first 61 belong to class 1, the remaining 28 to class 2.

The first subset used corresponds to the overall data set with 54675 probes.

The second subset counted 677 differentially expressed genes selected using Significance Analysis for Microarrays (15) and false discovery rate 5%.

The third subset was obtained performing a SAM selection (15) on the data. It consisted of 2870 genes. In order to accomplish such a selection, we used the *sam* function of Bioconductor (16) performed with 100 iterations and α=2.5%. No correction for multiple testing was used, so to gain a comparable number of probes to that of the group afterwards obtained.

Probes of the fourth subset were extracted using MAD and Wilcoxon test. MAD (Median Absolute Deviation) was computed on both classes. Five samples from each class were then randomly chosen for 100 times, to have an estimate of both class MAD values. Assuming samples of the two groups to be independent and to belong to two different populations with the same but unknown distribution, and the same standard deviation, Wilcoxon test was performed. Using a significance level given by Bonferroni correction $\alpha=0.025/N$, 2320 genes were selected.
The last subset is given by the union of the probes of the third and fourth data set, and contained 4698 probes.
That said, each group is referred to as in Table 4.

	Subset	Selection mode	# of genes
1	CLL_89_tot	no selection	54675
2	CLL_89_sam_fdr	sam fdr=5%	677
3	CLL_89_sam	sam α=2.5%	2870
4	CLL_89_mad	mad + wilcoxon	2320
5	CLL_89_union	sam $\cup$ mad	4698

Table 4

4.2. Result evaluation

As reported by Peng Qiu et al., and as had become clear during the thesis, SPD is very sensible to the input data, meaning that, depending on the genes provided, results may greatly vary. Particularly, we needed a criterion to evaluate how to set the user defined parameters in SPD. We focused mainly on the standard deviation threshold for gene filtering, because of the strong dependence of results on such parameter. To find a rational decision rule, we tested SPD on each group using different threshold values, and then attempted to extract those giving best results. Standard deviation threshold values used ranged between 0 and the maximum value still producing at least two meaningful modules, with a pitch of 0.1. SPD actually gave an error when building the progression similarity matrix if the modules to compare were less than two. For subset CLL_89_tot, the initial standard deviation value was not 0 but 0.5, due to very long running times to complete the computation of the clustering step and the comparison between modules.

In Figure 4 it is shown, as an example, the output of SPD. It is the results obtained from subset CLL_89_sam for standard deviation threshold value equal to 0.9. Blue dots represent patients with positive prognosis, whereas green diamonds represent patients with poor prognosis.

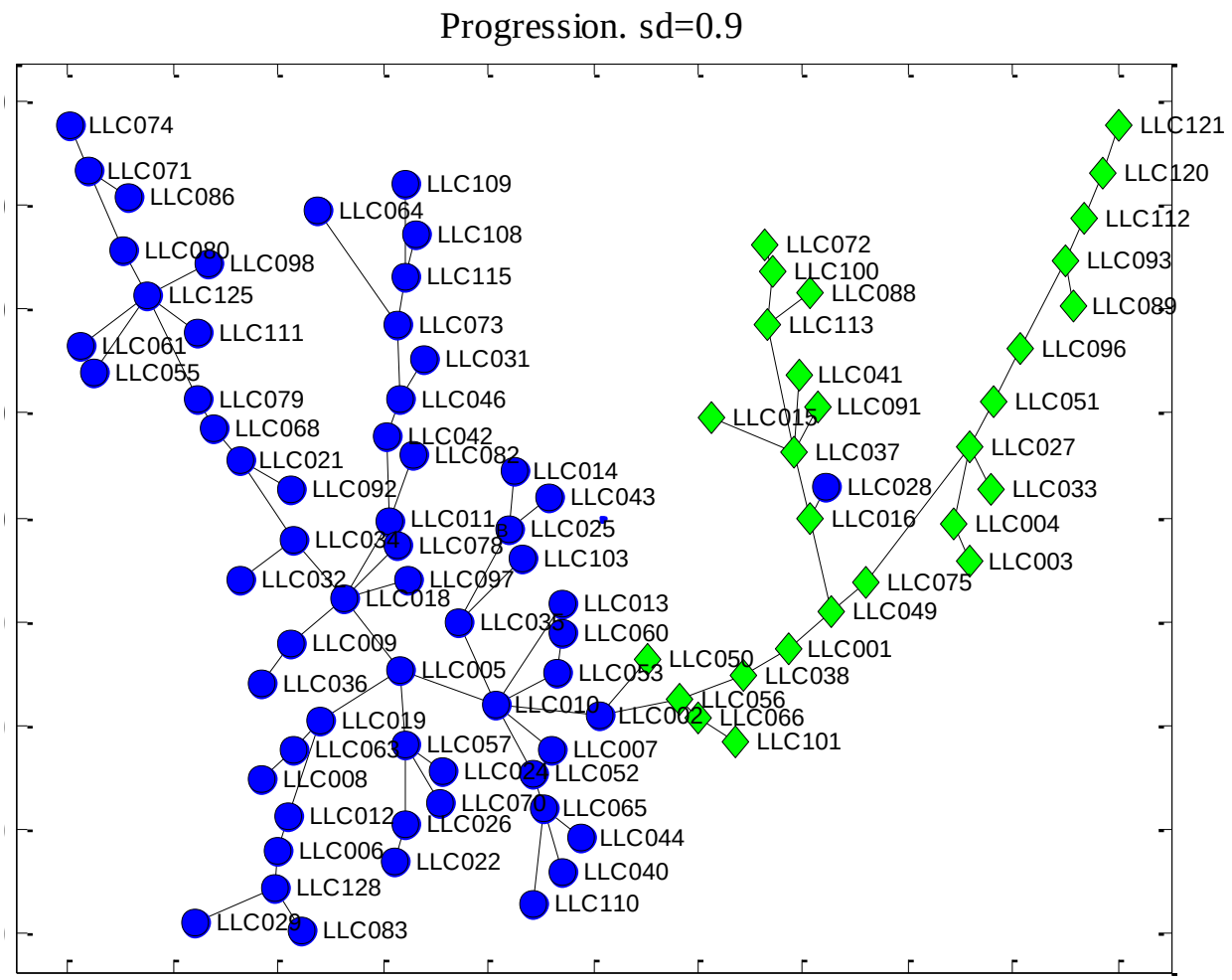

Figure 4

To evaluate the reliability of the results obtained, we checked whether the subdivision in the two prognostic classes was correctly or not recovered. The measure of correctness was made by manually counting how many samples were misclassified. And by dividing it by 89.

Once we had the relative error trend and its average, we restricted the range of standard deviation values to those ones giving an “acceptable” error. As a general rule, we considered an error to be “acceptable” if it was below the average.

Based on the chosen range of standard deviation threshold, we computed the pairwise distances between all the progressions obtained using the corresponding adjacency matrices. Given two

adjacency matrices A and B, defined as in section 3.2.5, and $S = |A - B|$, the distance was computed as the following score:

$$s = \frac{\sum_{S_{ij} \neq 0} S}{N \cdot (N-1)}$$

Element S_{ij} of matrix S is equal to 1 if and only if in one of the progressions there is an edge connecting two samples that is missing in the other one. Thus, the greater s is, the greater is the distance between the two progressions. The distance s was plotted versus $\Delta sd = |sd_A - sd_B|$ using boxplots.

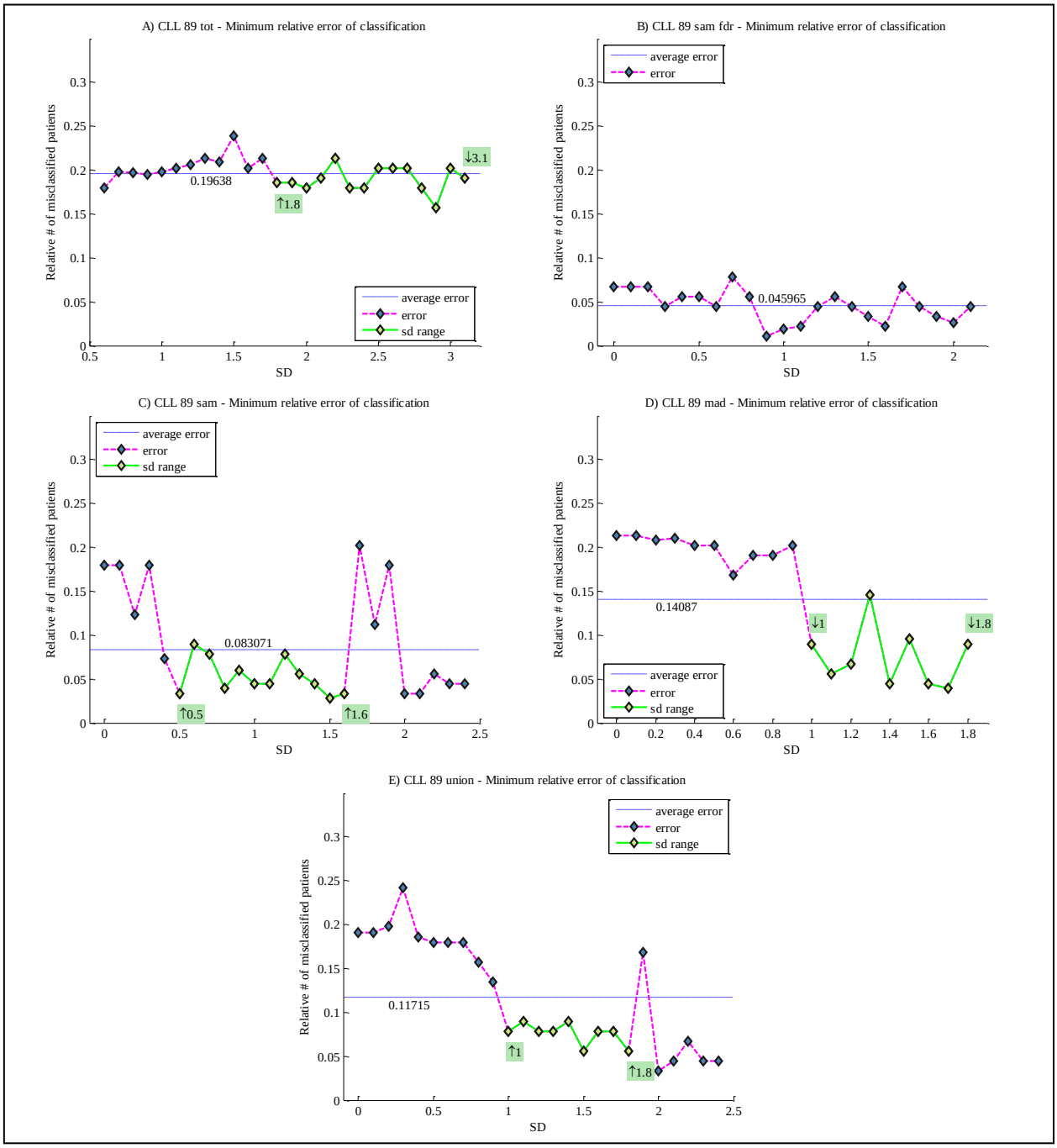

Figure 5

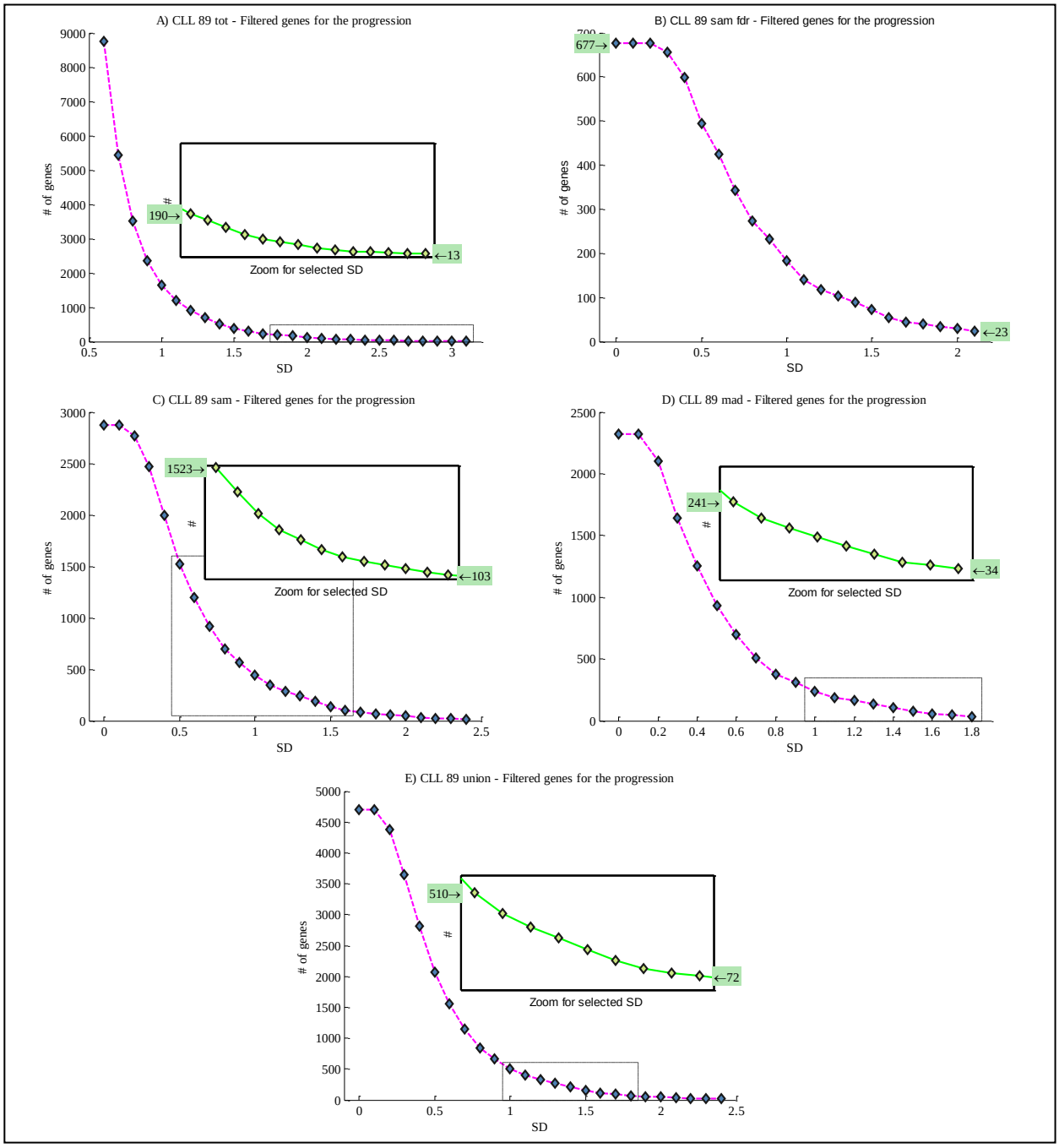
A) CLL 89 tot - Filtered genes for the progression
# of genes
SD
#
190→
←13
Zoom for selected SD
B) CLL 89 sam fdr - Filtered genes for the progression
677→
←23
C) CLL 89 sam - Filtered genes for the progression
1523→
←103
D) CLL 89 mad - Filtered genes for the progression
241→
←34
E) CLL 89 union - Filtered genes for the progression
510→
←72

Figure 6

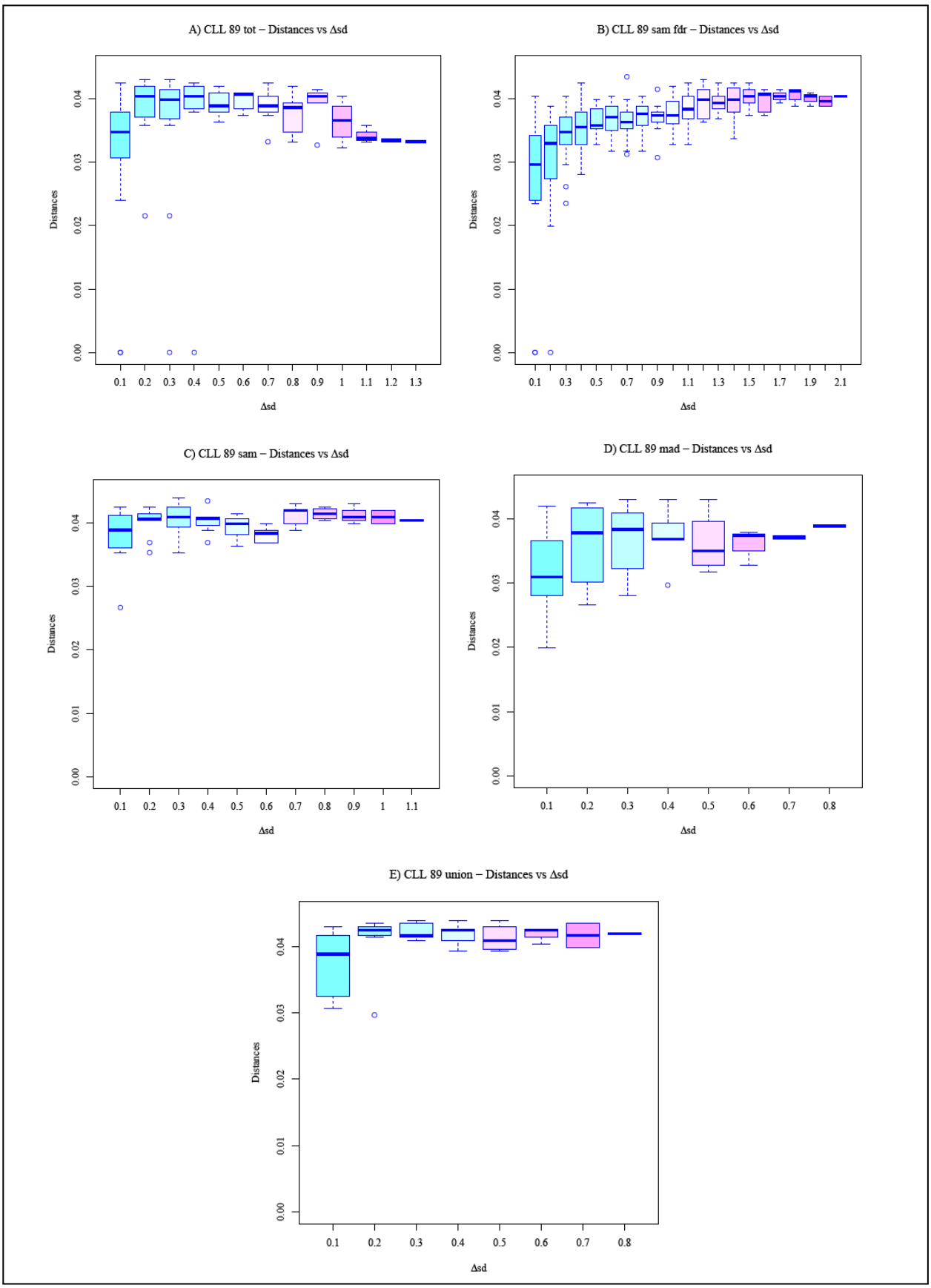
A) CLL 89 tot – Distances vs Δsd
B) CLL 89 sam fdr – Distances vs Δsd
C) CLL 89 sam – Distances vs Δsd
D) CLL 89 mad – Distances vs Δsd
E) CLL 89 union – Distances vs Δsd
Distances
Δsd

Figure 7

For each subset and for each value of standard deviation threshold we obtained a progression of the same kind as the one shown in Figure 4. For each output given by SPD, we computed and plotted the relative error of classification and the number of filtered genes versus standard deviation threshold values (Figure 5 and Figure 6). We got two plots for each subset of genes. Based on these figures, for each subset, we chose a range of standard deviation threshold. As a criterion to make such choice, we looked at the relative error trend, and we selected a range of values for which the error was below the average for that subset. Furthermore, we highlighted the number of genes filtered for the selected range of standard deviation threshold (Figure 6), and we computed and plotted the distances between progressions versus the difference of standard deviation threshold (Figure 7).

Figure 5 shows the behavior of the relative classification error versus standard deviation threshold, for each subset of genes.

It is evident from the average error trends that gene selection is essential. Indeed, when using all the genes of the microarray to find a progression, SPD produced results with high misclassification error, regardless of the standard deviation threshold (Figure 5-A). This experimental evidence suggests that gene expression values for the whole microarray include too much noise. Such noise exceeds the information content peculiar for that pathology. We can also see in Figure 6 that the number of filtered genes that produce good results varies a lot.
In Figure 5-B the average classification error is the lowest obtained, meaning that in the progressions most of the patients were recognized as belonging to their actual prognostic class. For this reason we decided not to select any restricted range of standard deviation threshold values.
From the trend of classification error for CLL_89_sam subset (Figure 5-C) is clear which is the best range of standard deviation threshold values, and the average error is the second smallest obtained. The correspondent number of filtered genes is the highest, ranging from 1523 to 103.
The third method of gene selection that we computed, producing subset CLL_89_mad is shown in Figure 5-D, was the second worst result obtained. Indeed, the average relative error is approximately 14%. This result had also a negative influence on subset CLL_89_union, since it was given by the union of subset CLL_89_sam with subset CLL_89_mad. It seems fair to expect a result that is a combination of the two previous ones and, as a matter of facts, it is. In Figure 5-E we can see that the error trend has common characteristics with trends in Figure 5-C and D.

Moreover, the average error is about the average of the previous two averages, and so it is the number of genes for the selected range of standard deviation threshold.
Concerning the distances between progressions, from Figure 7 we can see how different progressions are from each others. Even for a small variation in standard deviation threshold, resulting progressions presented heterogeneous structures.

4.3. Conclusions

As expected, using rough data does not produce reliable results, as only a small fraction of genes is not specifically related to the disease. The greatest part of genes prevents SPD from recovering a progression, increasing the noise that exceeds the informative content. Thus, a selection of genes is absolutely necessary.
Among the methods we used to reduce the number of genes to extract the most meaningful ones, SAM selection with fdr=5% (CLL_89_sam_fdr) and SAM selection with α=2.5% (CLL_89_sam) gave the best results. SPD was able to classify subjects in the two prognostic classes, with low misclassification error.
In particular, for subset CLL_89_sam, the average error was 8.3% and, for standard deviation threshold values ranging from 0.5 to 1.6, the error was below the average. The corresponding number of genes used to produce the progressions ranged from 1523 to 103. Subset CLL_89_sam_fdr, on the other hand, gave good results for standard deviation threshold values ranging from 0 to 2.1 and a number of genes ranging from 677 to only 23. The average error of classification was 4.6%, being the lowest obtained.

The worst results after gene selection was given by subset CLL_89_mad.
Knowing that results are strongly affected by the initial gene selection, as a general rule, we could suggest to use a standard deviation threshold value, chosen among those ones giving acceptable results, that is as conservative as possible in regard to the number of genes. As a consequence, it seems reasonable to make several trials, and choose suitable parameter values by visual inspection of the results.

Section 5

APPLICATION TO CHRONIC LYMPHOCYTIC LEUKEMIA

After evaluating the behavior of SPD on different inputs and several settings of standard deviation threshold parameter, we applied it to another two data sets associated to chronic lymphocytic leukemia and Waldenström's macroglobulinemia.

5.1. Gene selection

Given the results previously obtained, for this part of the thesis, gene selection was performed by using significance analysis of microarrays (15) with false discovering rate 5%. Thus, from an initial amount of 54675 genes, only 4374 were selected. The number of probes, i.e. of patients, was 114 split in the following groups:

1. 62 patients with positive prognosis (M-IgV_H and ZAP70$^-$) referred to as "Positive";
2. 28 patients with poor prognosis (UM-IgV_H and ZAP70$^+$) referred to as "Negative";
3. 23 patients with undefined prognosis (UM-IgV_H and ZAP70$^-$, or M-IgV_H and ZAP70$^+$) referred to as "NC";
4. 1 patient with unknown prognosis referred to as "NA".

This subset is referred to as CLL_114.

CLL_114		
Class name	# of patients	Prognosis
Positive	62	Positive
Negative	28	Poor
NC	23	Undefined
NA	1	Non available

Table 5

5.2. SPD results

Basing on previous results and being aware that SPD is extremely sensible to the input, when we applied it to subset CLL_114, we used values of standard deviation threshold ranging from 0.5 to 1.6. For each of them, we evaluated the error of classification as follows. As in this situation there were three classes to be distinguished, we exploited the fact that class "Negative" was clearly isolated from class "Positive" and class "NC", for every value of standard deviation threshold. Thus, we assumed patients belonging to class "Positive" and class "NC" as being part of the same group, so that the error of classification could be computed as described in 4.2 for CLL_89 subsets. As shown in Figure 8, relative classification error is much more regular compared to previous results, and close to the average value, which is 5.75%.

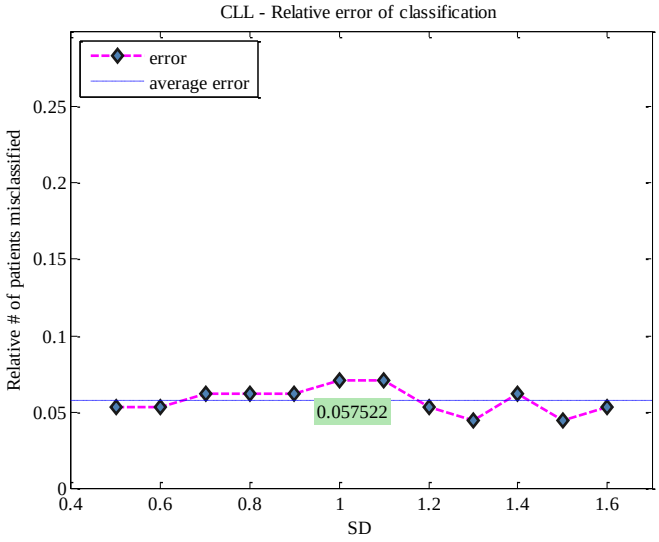

Figure 8

We show in Figure 9 one of the progression produced by SPD, to summarize the results obtained for this data set.

SPD cuts off poor prognosis samples and groups together patients with a good or undefined prognosis. According to this result, patients presenting prognostic marker values not following the standard for prognostic classification, should be considered and treated as those ones with a positive prognosis.

It is interesting to point out that patient "LLC043" precedes "LLC043.2", which is actually the same patient who underwent RNA exam twice at different times. This patient has been stable since diagnosis, and is not being treated as presents a positive prognosis. Only one time, for sd=1.0, they were both misclassified, but still linked to each other.

A closer look to patients in class “NC” revealed that their association to patients with positive prognosis was due to the mutational status of IgV_H. Indeed, IgV_H is a more relevant prognostic biomarker, rather than ZAP70 expression.

Another result to highlight is about patient “LLC160”. SPD classified it with classes “Positive” and “NC” patients for all standard deviation threshold values but for sd=1.4. Furthermore, to such value corresponds a relative classification error greater than the average. Hence, it could be considered to have a positive prognosis. When that was the case, “LLC160” position along the progression was either close to class “Negative” misclassified patients, or among class “Positive” and “NC” patients.

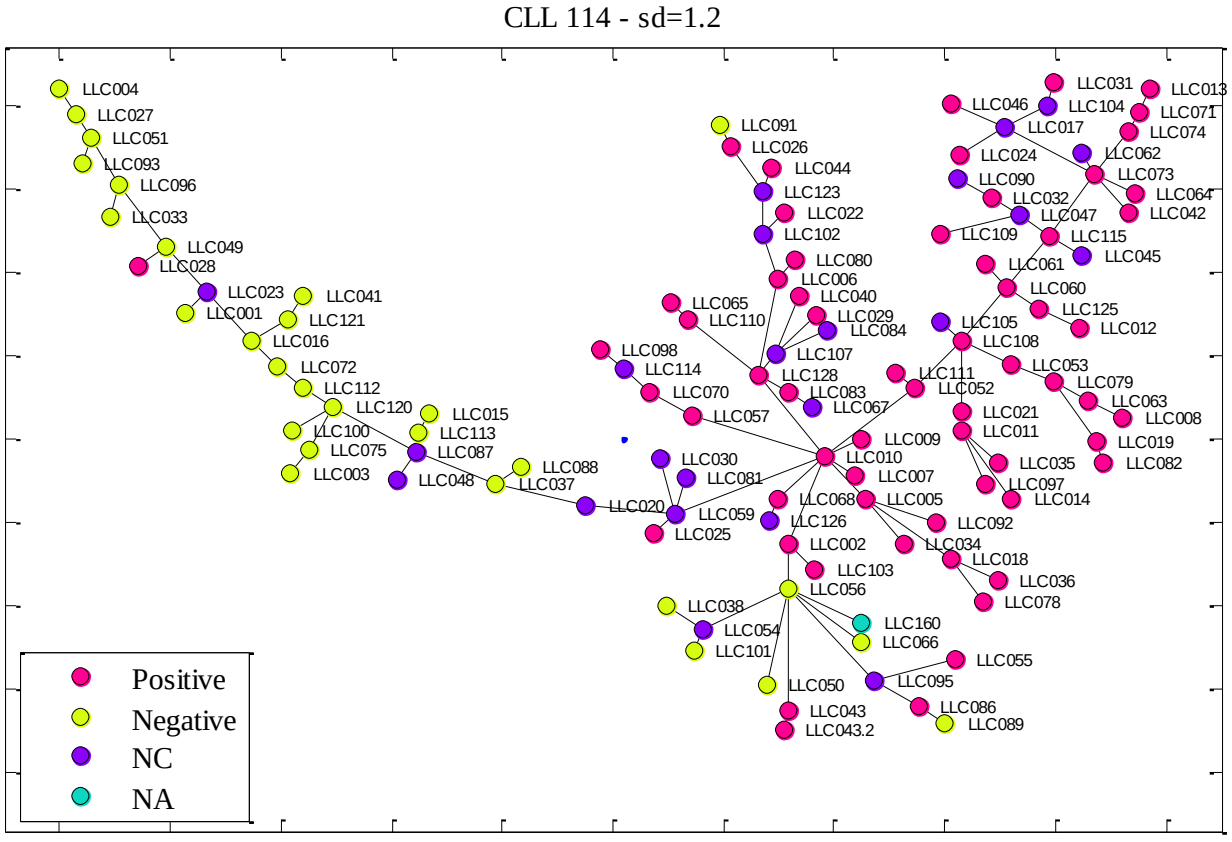

Figure 9

We chose, as the most representative progression, the one given by sd=1.2, because for this value the relative misclassification error is below average and subjects “LLC043”, “LLC043.2” and “LLC160” position reflects the overall outcomes previously depicted.

We extracted the gene lists actually used to build each progression and put them together in two different ways: we evaluated their union and intersection. For this data set and for the standard deviation threshold values considered, the union included 213 genes, as represented in Table 6, whereas the intersection was empty. Among all of those genes, the ones highlighted in the table have already been pointed out in several studies as being related to IgV_H mutational status, and could be considered eligible candidate as additional prognostic biomarkers (8).

For the sake of completeness, we performed a functional annotation clustering on genes reported in Table 6. We used the "functional annotation clustering" tool available on DAVID Bioinformatics Resources (17), (18) to cluster together genes with similar annotation. Results are shown in Table 11 in Section 9.

Table 6

PROBE ID	GENE ID	GENE NAME
205978_at	KL	klotho
210401_at	P2RX1	purinergic receptor P2X, ligand-gated ion channel, 1
1554733_at	MGC24125	hypothetical protein MGC24125
227530_at	AKAP12	A kinase (PRKA) anchor protein 12
222453_at	cybrd1	cytochrome b reductase 1
227265_at	FGL2	fibrinogen-like 2
204254_s_at	VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor
230287_at	SGSM1	small G protein signaling modulator 1
214453_s_at	IFI44	interferon-induced protein 44
206181_at	SLAMF1	signaling lymphocytic activation molecule family member 1
1556209_at	CLEC2B	C-type lectin domain family 2, member B
231356_at	LOC100131014	hypothetical LOC100131014
215145_s_at	CNTNAP2	contactin associated protein-like 2
230578_at	ZNF471	zinc finger protein 471
209815_at	ptch1	patched homolog 1 (Drosophila)
232584_at	TSHZ2	teashirt zinc finger homeobox 2
1553196_a_at	FCRL3	Fc receptor-like 3
212446_s_at	LASS6	LAG1 homolog, ceramide synthase 6
238071_at	LCN10	lipocalin 10
204083_s_at	tpm2	tropomyosin 2 (beta)
212698_s_at	SEPT10	septin 10
230831_at	Frmd5	FERM domain containing 5
219255_x_at	Il17rb	interleukin 17 receptor B
229552_at	LOC283454	hypothetical protein LOC283454
202393_s_at	KLF10	Kruppel-like factor 10
212985_at	Apbb2	amyloid beta (A4) precursor protein-binding, family B, member 2
203030_s_at	Ptprn2	protein tyrosine phosphatase, receptor type, N polypeptide 2
211637_x_at	LOC100126583	hypothetical LOC100126583

224156_x_at	Il17rb	interleukin 17 receptor B
226425_at	CLIP4	CAP-GLY domain containing linker protein family, member 4
229344_x_at	RIMKLB	ribosomal modification protein rimK-like family member B
218418_s_at	KANK2	KN motif and ankyrin repeat domains 2
231303_at	NCRNA00158	non-protein coding RNA 158
206978_at	CCR2	chemokine (C-C motif) receptor 2
206100_at	CPM	carboxypeptidase M
204646_at	DPYD	dihydropyrimidine dehydrogenase
221802_s_at	KIAA1598	KIAA1598
244740_at	MGC9913	hypothetical protein MGC9913
223380_s_at	Lats2	LATS, large tumor suppressor, homolog 2 (Drosophila)
238983_at	nsun7	NOL1/NOP2/Sun domain family, member 7
219304_s_at	PDGFD	platelet derived growth factor D
225897_at	MARCKS	myristoylated alanine-rich protein kinase C substrate
215489_x_at	HOMER3	homer homolog 3 (Drosophila)
213566_at	RNASE6	ribonuclease, RNase A family, k6
204834_at	FGL2	fibrinogen-like 2
211474_s_at	serpinb6	serpin peptidase inhibitor, clade B (ovalbumin), member 6
203796_s_at	BCL7A	B-cell CLL/lymphoma 7A
211643_x_at	IGKV3D-15	immunoglobulin kappa variable 3D-15 (gene/pseudogene)
219300_s_at	CNTNAP2	contactin associated protein-like 2
236918_s_at	LRRC34	leucine rich repeat containing 34
204454_at	Ldoc1	leucine zipper, down-regulated in cancer 1
207120_at	ZNF667	zinc finger protein 667
234284_at	GNG8	guanine nucleotide binding protein (G protein), gamma 8
219738_s_at	PCDH9	protocadherin 9
226926_at	Dmkn	dermokine
203642_s_at	Cobll1	COBL-like 1
1560562_a_at	ZNF677	zinc finger protein 677
223620_at	GPR34	G protein-coupled receptor 34
225133_at	KLF3	Kruppel-like factor 3 (basic)
209674_at	CRY1	cryptochrome 1 (photolyase-like)
205414_s_at	RICH2	Rho-type GTPase-activating protein RICH2
228557_at	L3mbtl4	l(3)mbt-like 4 (Drosophila)
214720_x_at	SEPT10	septin 10
212442_s_at	LASS6	LAG1 homolog, ceramide synthase 6
216491_x_at	IGHV3-11	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
216491_x_at	IGHV3-7	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
216491_x_at	IGHG3	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus

216491_x_at	Ighg1	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
216491_x_at	IGH@	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
216491_x_at	IGHM	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
216491_x_at	ighv4-31	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
214032_at	zap70	zeta-chain (TCR) associated protein kinase 70kDa
202241_at	TRIB1	tribbles homolog 1 (Drosophila)
228855_at	NUDT7	nudix (nucleoside diphosphate linked moiety X)-type motif 7
209854_s_at	KLK2	kallikrein-related peptidase 2
201670_s_at	MARCKS	myristoylated alanine-rich protein kinase C substrate
225864_at	FAM84B	family with sequence similarity 84, member B
210102_at	vwa5a	von Willebrand factor A domain containing 5A
244741_s_at	MGC9913	hypothetical protein MGC9913
242064_at	sdk2	sidekick homolog 2 (chicken)
201540_at	fhl1	four and a half LIM domains 1
227013_at	Lats2	LATS, large tumor suppressor, homolog 2 (Drosophila)
220066_at	NOD2	nucleotide-binding oligomerization domain containing 2
228297_at	cnn3	calponin 3, acidic
211640_x_at	IGHV1-69	immunoglobulin heavy variable 1-69; similar to hCG1773549
211640_x_at	LOC100133862	immunoglobulin heavy variable 1-69; similar to hCG1773549
219302_s_at	CNTNAP2	contactin associated protein-like 2
232821_at	GTSF1L	gametocyte specific factor 1-like
204334_at	KLF7	Kruppel-like factor 7 (ubiquitous)
235570_at	RBMS3	RNA binding motif, single stranded interacting protein
203548_s_at	Lpl	lipoprotein lipase
238870_at	KCNK9	potassium channel, subfamily K, member 9
226485_at	VSIG10	hypothetical protein FLJ20674
228494_at	ppp1r9a	protein phosphatase 1, regulatory (inhibitor) subunit 9A
236894_at	L1TD1	LINE-1 type transposase domain containing 1
227529_s_at	AKAP12	A kinase (PRKA) anchor protein 12
1552736_a_at	NETO1	neuropilin (NRP) and tolloid (TLL)-like 1
206983_at	CCR6	cyclin L2; chemokine (C-C motif) receptor 6
206983_at	Ccnl2	cyclin L2; chemokine (C-C motif) receptor 6
212190_at	SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2
236600_at	spg20	spastic paraplegia 20 (Troyer syndrome)
226517_at	BCAT1	branched chain aminotransferase 1, cytosolic
202555_s_at	MYLK	myosin light chain kinase
231358_at	mro	maestro

235616_at	TSHZ2	teashirt zinc finger homeobox 2
200897_s_at	palld	palladin, cytoskeletal associated protein
223595_at	TMEM133	transmembrane protein 133
203549_s_at	Lpl	lipoprotein lipase
203705_s_at	FZD7	frizzled homolog 7 (Drosophila)
232383_at	TFEC	transcription factor EC
205771_s_at	AKAP7	A kinase (PRKA) anchor protein 7
210612_s_at	SYNJ2	synaptojanin 2
202599_s_at	NRIP1	nuclear receptor interacting protein 1
203641_s_at	Cobll1	COBL-like 1
219737_s_at	PCDH9	protocadherin 9
235743_at	SNED1	sushi, nidogen and EGF-like domains 1
243375_at	GRIK1	glutamate receptor, ionotropic, kainate 1
210644_s_at	Lair1	leukocyte-associated immunoglobulin-like receptor 1
204731_at	Tgfbr3	transforming growth factor, beta receptor III
219496_at	ANKRD57	ankyrin repeat domain 57
205992_s_at	IL15	interleukin 15
209732_at	CLEC2B	C-type lectin domain family 2, member B
204072_s_at	FRY	furry homolog (Drosophila)
226625_at	Tgfbr3	transforming growth factor, beta receptor III
215767_at	ZNF804A	zinc finger protein 804A
213714_at	CACNB2	calcium channel, voltage-dependent, beta 2 subunit
235800_at	ENO4	chromosome 10 open reading frame 134
221261_x_at	MAGED4B	melanoma antigen family D, 4B; melanoma antigen family D, 4
221261_x_at	MAGED4	melanoma antigen family D, 4B; melanoma antigen family D, 4
218613_at	PSD3	pleckstrin and Sec7 domain containing 3
213906_at	mybl1	v-myb myeloblastosis viral oncogene homolog (avian)-like 1
206865_at	HRK	harakiri, BCL2 interacting protein (contains only BH3 domain)
239246_at	FARP1	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
212655_at	ZCCHC14	zinc finger, CCHC domain containing 14
227792_at	Itpripl2	inositol 1,4,5-triphosphate receptor interacting protein-like 2
213093_at	Prkca	protein kinase C, alpha
212503_s_at	dip2c	DIP2 disco-interacting protein 2 homolog C (Drosophila)
236635_at	ZNF667	zinc finger protein 667
227034_at	ANKRD57	ankyrin repeat domain 57
227810_at	ZNF558	zinc finger protein 558
229347_at	LOC729506	hypothetical LOC729506
244521_at	TSHZ2	teashirt zinc finger homeobox 2
214452_at	BCAT1	branched chain aminotransferase 1, cytosolic
206115_at	EGR3	early growth response 3
1562713_a_at	NETO1	neuropilin (NRP) and tolloid (TLL)-like 1
230793_at	Lrrc16a	leucine rich repeat containing 16A
223535_at	NUDT12	nudix (nucleoside diphosphate linked moiety X)-type motif 12
214953_s_at	APP	amyloid beta (A4) precursor protein
1556839_s_at	SPTBN5	spectrin, beta, non-erythrocytic 5
201669_s_at	MARCKS	myristoylated alanine-rich protein kinase C substrate

243940_at	TSHZ2	teashirt zinc finger homeobox 2
212526_at	spg20	spastic paraplegia 20 (Troyer syndrome)
233985_x_at	ppp1r9a	protein phosphatase 1, regulatory (inhibitor) subunit 9A
224361_s_at	Il17rb	interleukin 17 receptor B
201876_at	PON2	paraoxonase 2
238447_at	RBMS3	RNA binding motif, single stranded interacting protein
1569346_a_at	P2RX1	purinergic receptor P2X, ligand-gated ion channel, 1
225285_at	BCAT1	branched chain aminotransferase 1, cytosolic
224823_at	MYLK	myosin light chain kinase
219841_at	AICDA	activation-induced cytidine deaminase
211634_x_at	IGHV1-69	immunoglobulin heavy variable 1-69; similar to hCG1773549
211634_x_at	LOC100133862	immunoglobulin heavy variable 1-69; similar to hCG1773549
216620_s_at	arhgef10	Rho guanine nucleotide exchange factor (GEF) 10
203029_s_at	Ptprn2	protein tyrosine phosphatase, receptor type, N polypeptide 2
202342_s_at	trim2	tripartite motif-containing 2
204647_at	HOMER3	homer homolog 3 (Drosophila)
211635_x_at	IGHV1OR15-9	V-set and immunoglobulin domain containing 7; immunoglobulin heavy variable 1/OR15-5 pseudogene; immunoglobulin heavy variable 1/OR15-9 (non-functional)
211635_x_at	VSIG7	V-set and immunoglobulin domain containing 7; immunoglobulin heavy variable 1/OR15-5 pseudogene; immunoglobulin heavy variable 1/OR15-9 (non-functional)
211635_x_at	IGHV1OR21-1	V-set and immunoglobulin domain containing 7; immunoglobulin heavy variable 1/OR15-5 pseudogene; immunoglobulin heavy variable 1/OR15-9 (non-functional)
241278_at	FCRL3	Fc receptor-like 3
226164_x_at	RIMKLB	ribosomal modification protein rimK-like family member B
225330_at	IGF1R	insulin-like growth factor 1 receptor
228532_at	C1orf162	chromosome 1 open reading frame 162
207245_at	UGT2B17	UDP glucuronosyltransferase 2 family, polypeptide B17
1560225_at	cnr1	cannabinoid receptor 1 (brain)
205419_at	Gpr183	G protein-coupled receptor 183
238577_s_at	TSHZ2	teashirt zinc finger homeobox 2
228033_at	E2F7	E2F transcription factor 7
217371_s_at	IL15	interleukin 15
221337_s_at	ADAM29	ADAM metallopeptidase domain 29
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3
223696_at	arsD	arylsulfatase D
203795_s_at	BCL7A	B-cell CLL/lymphoma 7A
204439_at	IFI44L	interferon-induced protein 44-like
203881_s_at	dmd	dystrophin
213436_at	cnr1	cannabinoid receptor 1 (brain)
225140_at	KLF3	Kruppel-like factor 3 (basic)
226247_at	plekha1	pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 1
238512_at	CAPZA1	capping protein (actin filament) muscle Z-line, alpha 1
202600_s_at	NRIP1	nuclear receptor interacting protein 1
216541_x_at	IGHV1-69	immunoglobulin heavy variable 1-69; similar to hCG1773549
216541_x_at	LOC100133862	immunoglobulin heavy variable 1-69; similar to hCG1773549
226846_at	phyhd1	phytanoyl-CoA dioxygenase domain containing 1

1569345_at	P2RX1	purinergic receptor P2X, ligand-gated ion channel, 1
224499_s_at	AICDA	activation-induced cytidine deaminase
238778_at	MPP7	membrane protein, palmitoylated 7 (MAGUK p55 subfamily member 7)
221704_s_at	vps37b	vacuolar protein sorting 37 homolog B (S. cerevisiae)
230673_at	PKHD1L1	polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1
200602_at	APP	amyloid beta (A4) precursor protein
231093_at	FCRL3	Fc receptor-like 3
232820_s_at	GTSF1L	gametocyte specific factor 1-like
228737_at	TOX2	TOX high mobility group box family member 2
219955_at	L1TD1	LINE-1 type transposase domain containing 1
211633_x_at	IGHV3-11	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	IGHV3-7	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	IGHG3	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	Ighg1	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	IGH@	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	IGHM	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
211633_x_at	ighv4-31	immunoglobulin heavy constant gamma 1 (G1m marker); immunoglobulin heavy constant mu; immunoglobulin heavy variable 3-7; immunoglobulin heavy constant gamma 3 (G3m marker); immunoglobulin heavy variable 3-11 (gene/pseudogene); immunoglobulin heavy variable 4-31; immunoglobulin heavy locus
222457_s_at	LIMA1	LIM domain and actin binding 1
214039_s_at	Laptm4b	lysosomal protein transmembrane 4 beta
203695_s_at	Dfna5	deafness, autosomal dominant 5
204255_s_at	VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor
201911_s_at	FARP1	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
206486_at	LAG3	lymphocyte-activation gene 3
229598_at	Cobll1	COBL-like 1
222258_s_at	SH3BP4	SH3-domain binding protein 4
210517_s_at	AKAP12	A kinase (PRKA) anchor protein 12
238919_at	PCDH9	protocadherin 9
228974_at	ZNF677	zinc finger protein 677
221088_s_at	ppp1r9a	protein phosphatase 1, regulatory (inhibitor) subunit 9A

213056_at	FRMD4B	FERM domain containing 4B
227379_at	MBOAT1	membrane bound O-acyltransferase domain containing 1
203706_s_at	FZD7	frizzled homolog 7 (Drosophila)

Section 6

APPLICATION TO WALDENSTRÖM'S MACROGLOBULINEMIA AND IgM MGUS

The second application that with considered in this thesis regarded IgM monoclonal gammopathy of undetermined significance and Waldensström's macroglobulinemia. In particular we were concerned by the possible evolution of the former one in the latter one.

6.1. Gene selection

As explained in 2.2, this data set consisted in 97 probes extracted from different cell types. The first problem we had to deal with was a strong batch effect on microarray data. Batch effect, that is a non biological experimental variation, is pretty common in microarray experiments and it makes it inappropriate to combine data sets without adjusting for it. We used a function which exploits an empirical Bayesian framework (19). Named function is *ComBat* in the *sva* R package of Bioconductor (20).

The next step consisted in performing a SAM selection (15) only on genes belonging to those microarrays extracted from antigen CD19 positive cells. By using a false discorvery rate of 5%, we obtained 750 genes for each of the 38 samples. These samples, corresponding to as many individuals, presented two different diagnosis:

1. IgM MGUS: 11 patients (referred to as "MGUS");
2. WM: 27 patients (referred to as "WM").

This subset is referred to as WM_MGUS_38.

WM_MGUS_38		
Class name	# of patients	Diagnosis
MGUS	11	IgM MGUS
WM 2	27	WM

Table 7

6.2. SPD results

We had to make some attempts before having good results according to the trend of the relative classification error and hence the ability to distinguish between the two pathological conditions. At first, we divided the data set in 3 subsets of probes. Each subset consisted of the same number of genes, i.e. 54675 as no gene selection was performed, and samples were grouped together according to the cell type: $CD19^{+}$, $CD138^{+}$, NEG. SPD wasn't able to retrieve a progression representative of the different diagnosis for the two latter subsets, and results on the former subset were affected by the large number of genes, as expected. We then performed gene selection for each group, with the method explained in section 6.1. Again results for $CD138^{+}$ and NEG groups were not outstanding, meaning that the information content is in $CD19^{+}$ cells.

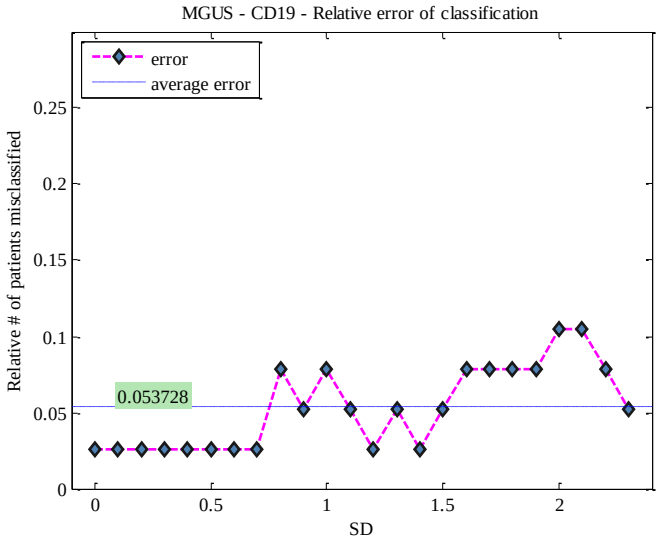

Figure 10

The relative classification error, shown in Figure 10, was computed by counting the number of sample swaps needed to have the right classification, and dividing by the total number of samples, namely 38.

Results gained on subset WM_MGUS_38 were satisfactory, as the trend of the relative classification error showed: the average relative error is approximately 5.3%. SPD was capable of distinguishing between IgM MGUS and WM patients. For a restricted range of standard deviation threshold values the number of filtered genes was constant, as consequently was the correspondent error.

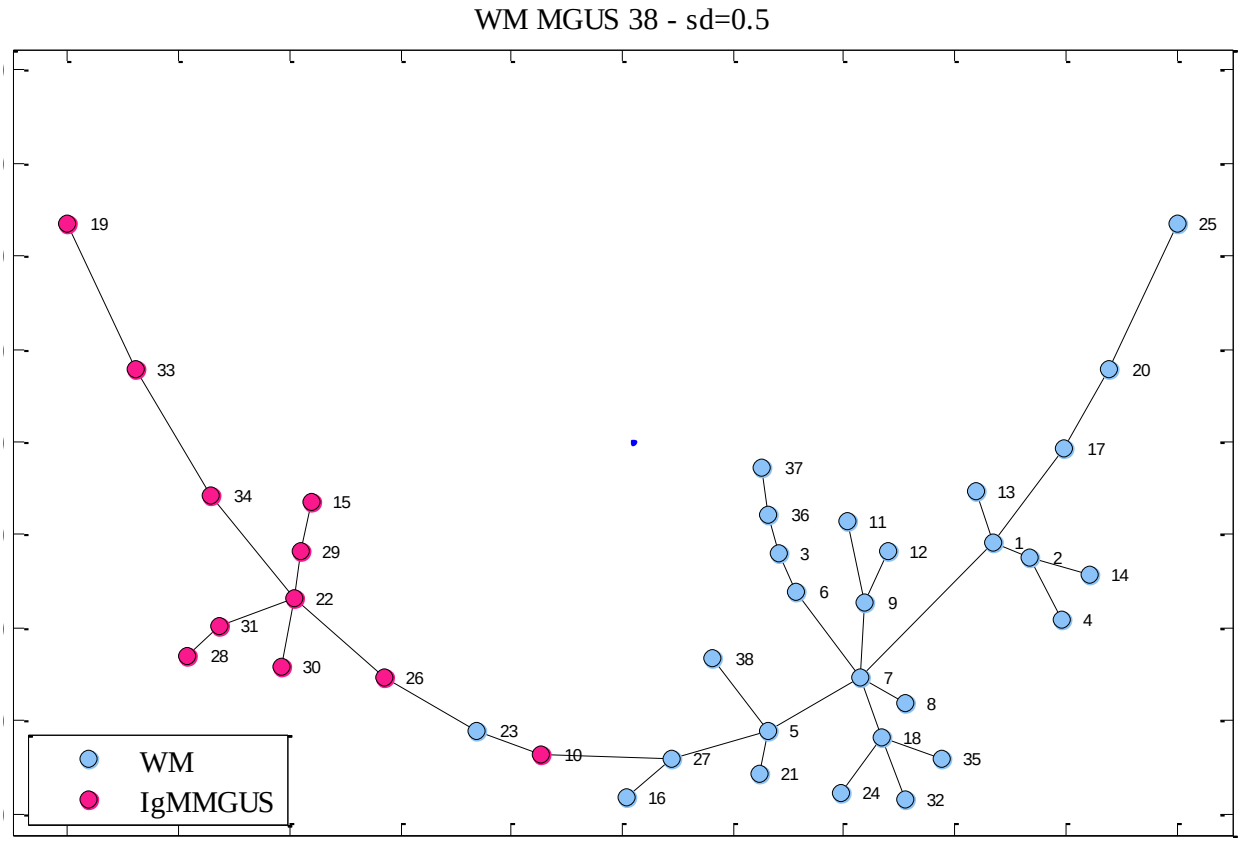

Figure 11

In Figure 11, we show one of the progressions made by SPD for standard deviation threshold set to 0.5.

For this subset of genes, other labels for each probe were available: WM label, bone marrow infiltration percentage, sex and age, as reported in Table 10. For each progression given by SPD, we checked if patients were classified according to some labels other than diagnosis. None of the progressions reflected the partition of patients according to any other labels.

Table 8 - Correspondence between SPD samples and microarray

#	/	Probe Name
1	/	2001_02_22_WM03_CD19_U133_PLUS2
2	/	2010_02_16_WM02_CD19
3	/	2010_03_15_WM01_CD19_U133_PLUS2
4	/	2010_03_25_WM04_CD19_U133_PLUS2
5	/	2010_03_25_WM06_CD19_U133_PLUS2
6	/	2010_03_31_WM07_CD19
7	/	2010_03_31_WM08_CD19
8	/	2010_04_08_WM09_CD19_PLUS2
9	/	2010_05_07_WM13_19_U133_PLUS2
10	/	2010_05_11_WM12_CD19_U133_PLUS2
11	/	2010_08_27_WM18_CD19_U133_PLUS2
12	/	2010_08_27_WM19_CD19_U133_PLUS2
13	/	2010_08_31_WM15_CD19_U133_PLUS2
14	/	2010_09_03_WM21_CD19_U133_PLUS2
15	/	2011_02_18_WM22_CD19_U133_PLUS2
16	/	2011_02_24_WM23_CD19_U133_PLUS2
17	/	2011_03_01_MGUS02_CD19_U133_PLUS2
18	/	2011_03_02_MGUS03_CD19_U133_PLUS2
19	/	2011_05_17_WM24PV_CD19_U133_PLUS2
20	/	2011_05_20_WM25PV_CD19_U133_PLUS2
21	/	2011_05_20_WM31_CD19_U133_PLUS2
22	/	2011_06_09_WM26_CD19_U133_PLUS2
23	/	2011_06_09_WM27_CD19_U133_PLUS2
24	/	2011_06_17_WM28_CD19_U133_PLUS2
25	/	2011_09_20_WM35_CD19_U133_PLUS2
26	/	2011_10_27_MGUS16_CD19_U133_PLUS2
27	/	2012_01_11_WM05_CD19_U133PLUS2
28	/	2012_01_18_MGUS22_CD19_U133PLUS2
29	/	2012_04_11_MGUS24_CD19
30	/	2012_04_11_MGUS25_CD19
31	/	2012_04_11_MGUS28_CD19
32	/	2012_06_21_WM37_CD19
33	/	2012_06_29_MGUS17_CD19_2
34	/	2012_06_29_MGUS21_CD19_2
35	/	2012_07_04_WM39_CD19
36	/	2012_07_04_WM41_CD19
37	/	2012_07_04_WM42_CD19
38	/	2012_07_04_WM43_CD19

Table 9

PROBE ID	GENE ID	GENE NAME
232687_at	GPRIN3	GPRIN family member 3
1562153_a_at	PVT1	Pvt1 oncogene (non-protein coding)
224156_x_at	Il17rb	interleukin 17 receptor B
1562754_at	LOC339260	hypothetical protein LOC339260
230793_at	Lrrc16a	leucine rich repeat containing 16A
215767_at	ZNF804A	zinc finger protein 804A
1553333_at	C1orf161	chromosome 1 open reading frame 161
218309_at	Camk2n1	calcium/calmodulin-dependent protein kinase II inhibitor 1
203404_at	ARMCX2	armadillo repeat containing, X-linked 2
210550_s_at	RASGRF1	Ras protein-specific guanine nucleotide-releasing factor 1
1564077_at	GPRIN3	GPRIN family member 3
1556697_at	GPRIN3	GPRIN family member 3
203215_s_at	myo6	myosin VI
212935_at	MCF2L	MCF.2 cell line derived transforming sequence-like
224361_s_at	Il17rb	interleukin 17 receptor B
226408_at	TEAD2	TEA domain family member 2
209498_at	CEACAM1	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
228560_at	Cacna1d	calcium channel, voltage-dependent, L type, alpha 1D subunit
225757_s_at	CLMN	calmin (calponin-like, transmembrane)
229656_s_at	Eml6	echinoderm microtubule associated protein like 6
229147_at	rassf6	Ras association (RalGDS/AF-6) domain family member 6
227556_at	NME7	non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)
210640_s_at	GPER	G protein-coupled estrogen receptor 1
1560762_at	LOC285972	hypothetical protein LOC285972
242785_at	Eml6	echinoderm microtubule associated protein like 6
219255_x_at	Il17rb	interleukin 17 receptor B

As for subset CLL_114, we evaluated the union and the intersection of the genes used to build each progression given by standard deviation threshold values ranging from 0 to 1.5. Union included 739 genes, not reported, whereas the intersection included 31 genes, as represented in Table 9. None of them was found in literature as being related to the evolution of IgM MGUS in WM.

For the sake of completeness, we performed a functional annotation clustering on genes reported in Table 9, as in 5.2. Results are not reported as fdr values were too high, i.e. greater than 5%.

Section 7

DISCUSSION

Within the study of pathological conditions starting from the analysis of high-throughput data, the usual approach consists in using supervised classification algorithms. The initial information about classes, usually two, is given by clinicians and is often affected by uncertainty. As emerges from literature, the supervised approach frequently fails, for two main reasons: the definition of classes is a tough task itself, and each sample belongs to a different stage along the disease progression.

The rationale was to exploit an unsupervised approach to arrange samples, according to the progression state of a disease. To do so, we used a tool, Sample Progression Discovery (SPD), developed by Peng Qiu et al. (1), that, starting from gene expression data, seeks to retrieve sample position along a biologically meaningful progression, and extracts genes responsible for such progression.
We then applied SPD to two pathological conditions: chronic lymphocytic leukemia (CLL) and Waldenström's macroglobulinemia (WM). To the purpose, we had to evaluate the reliability of SPD results, first. Indeed, they depend on the input provided (gene selection) and SPD internal parameter setting.

We used a data set consisting of 89 patients, who were diagnosed CLL, with a well defined prognosis. We tried different methods of gene selection and, for each of the subsets we obtained, we varied parameter values. We then evaluated quality of the results relying on SPD ability to classify patients in the right prognostic class. This former analysis proved that gene selection is essential, the best method being Significance Analysis of Microarrays (15) with false discovery rate threshold 5%; it is also necessary to evaluate outputs with parameter values.

That said, we were enabled to focus on the application of SPD to hematological neoplasms.
We considered two distinct hematological neoplasms: chronic lymphocytic leukemia and Waldenström's macroglobulinemia. Both of them are B-cell malignancies, often compared in the

literature, but we studied them separately, as finding differences or similarities between them was beyond our aims (21).

One reason to choose CLL is because this is a widespread disease within the population. Indeed, it's one of the most common types of leukemia and, though it has been widely studied, the definition of prognostic classes is not well established yet. We then explored the possibility of using SPD to help us finding prognostic factors and supporting doctor's decisions.

As to Waldenström's macroglobulinemia, the reason that prompted us to focus on it was the following. The etiology of WM is largely unknown. Patients with IgM monoclonal gammopathy of undetermined significance (IgM MGUS) are at the greatest risk of developing WM compared to the general population, and, therefore, MGUS is considered a precursor of WM

For CLL, guidelines based on prognosis biomarkers have already been established, but not all patients behave according to such guidelines. It was proven that un-mutated IgV_H and ZAP70 positive expression is related to a poor prognosis, whereas mutated IgV_H and ZAP70 negative expression corresponds to a positive prognosis. We submitted to SPD a data set that included patients not respecting such classification, hence with undefined prognosis, and a single patient with non available prognosis. The result given by SPD was a progression in which poor prognosis patients were isolated from the others, who were mixed together. A possible explanation is that undefined prognosis patients should be treated the same way as those with a positive prognosis. We also evaluated genes used by SPD to get to the progression. Some of them have already been studied and found to be correlated to the disease progression (8). Further analysis is needed, since most of the genes involved in building the progression still have to be evaluated.

Nowadays, it is a well established notion that a subset of IgM monoclonal gammopathy of undetermined significance represents the precursor state of Waldenström's macroglobulinemia (21). We used a data set consisting of 38 patients, 11 of them diagnosed IgM MGUS and the other 27 suffering from WM. SPD was able to recover a progression characterized by patients being grouped together according to their diagnosis. Such result is only a starting point. First of all it confirms that the gene selection made was actually appropriate for SPD, since noise due to all gene expression values was reduced. On the other hand, genes extracted to build the progression have to be further studied, as it is not yet clear what is responsible for the evolution of IgM MGUS into WM.

Even if Sample Progression Discovery results are strongly affected by the way input is prepared, it can be a powerful tool. Indeed, it was capable of recognizing patient membership to their own class, according to gene expression profiles, thus providing at least an idea on the disease evolution and on how to handle uncertain situations. Moreover, it detects genes underlying progressions.

Acknowledgements

Vorrei ringraziare tutte le persone che, in diversa misura, mi hanno accompagnato durante questo percorso di studi e di vita, grazie ai quali è stata possibile la realizzazione di questo progetto.

Grazie ai miei genitori che mi hanno sostenuto, moralmente ed economicamente, senza i quali tutto ciò non sarebbe stato possibile. Con la speranza che sia stato ripagato anche il loro sforzo e che siano soddisfatti e orgogliosi.

Un sentito grazie agli amici conosciuti a Padova, con i quali ho condiviso gioie e dolori dell'esperienza universitaria. In particolare, ringrazio la Vivi per tutti gli anni passati in appartamento assieme, per esserci stata nei momenti bui, ma soprattutto, per esserci stata nei momenti felici. Grazie a Laura, per essere stata degna sostituta della Viviana e anche di più. E grazie a Stefano, che è sopravvissuto due anni circondato da ragazze in preda ad altalenanti stati d'animo.
Grazie a Giulio, fedele compagno di esami, e a tutti i compagni di corso, per le risate, per le chiacchierate, per le serate assieme, e per i loro immancabili commenti su ogni paio di orecchini o colore di smalto.

Un ringraziamento alla professoressa Di Camillo per avermi presa in tesi e avermi trasmesso tanti insegnamenti durante lo svolgimento del lavoro.
Grazie a Mark Bordovsky per avermi dedicato parte del suo tempo, consentendomi di imparare più di quanto abbia insegnato al suo corso. Grazie di cuore perché ha trasformato la mia esperienza in America, dandole un senso e arricchendola significativamente.

Grazie alla Michi, alla Silvia e a tutti coloro che non sono stati citati personalmente, ma a cui sono riconoscente in egual misura.

Section 8

REFERENCES

1. *Discovering Biological Progression Underlying Microarray Samples.* **Qiu, Peng, Gentles, Andrew J. and Plevritis, Sylvia K.** 4, 2011, PLoS Comput Biol, Vol. 7, pp. 1-11. doi:10.1371/journal.pcbi.1001123.

2. *Monoclonal Gammopathy of Undetermined Significance, Waldenstrom Macroglobulinemia, AL Amyloidosis, and related plasma cell disorders: diagnosis and treatment.* **Rajkumar, S. Vincent, Dispenzieri, Angela and Kyle, Robert A.** 2006. Symposium on oncology practice: hematological malignancies. pp. 693-703.

3. *Reconstructiong the temporal ordering of biological samples using microarray data.* **Magwene, Paul M., Lizardi, Paul and Kim, Junhyong.** 7, 2003, Bioinformatics, Vol. 19, pp. 842-850. doi:10.1093/bioinformatics/btg081.

4. *Control of hematopoietic differentiation: lack of specificity in signaling by cytokine receptors.* **Socolovsky, Merav, Lodish, Harvey F. and Daley, George Q.** 1998, Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 6573-6575.

5. *Chronic Lymphocytic Leukemia.* **Boelens, Jerina, et al.** 2009, Anticancer Research, Vol. 29, pp. 605-616.

6. *Chornic Lympocytic Leukemia.* **Chiorazzi, Nicholas, Rai, Kanti R. and Ferrarini, Manlio.** 2005, N Engl J Med, Vol. 352, pp. 804-815.

7. *ZAP-70 expression is associated with enhanced ability to respond to migratory and survival signals in B-cell chronic lymphocytic leukemia (B-CLL).* **Richardson, Sarah J., et al.** 2006, blood, Vol. 107, pp. 3584-3592. doi:10.1182/blood-2005-04-1718.

8. *Gene expression profiling idenfies ASRD as a new marker of disease progression and the sphingolipid metabolism as a potetial novel metabolism in chronic lympocitic leukemia.* **Trojani, Alessandra, et al.** 2011/2012, Cancer Biomarkers, Vol. 11, pp. 15-28. doi:10.3233/CBM-2012-0259.

9. *Lipoprotein lipase is differentially expressed in prognostic subsets of chronic lymphocytic leukemia but displays invariably low catalytical activity.* **Mansouri, Mahamoud, et al.** 3, 2010, Leukemia Research, Vol. 34, pp. 301-306. doi:10.1016/j.leukres.2009.07.032.

10. *Waldenstrom Macroglobulinemia: A Review of the Entity and Its Differential Diagnosis.* **Shaheen, Saad P., et al.** 1, 2012, Adv Anat Pathol, Vol. 19, pp. 11-27.

11. *An analysis of intra array repeats:the good, the bad, and the non informative.* **Elbez, Yedid, Farkash-Amar, Shlomit and Simon, Itamar.** 1, 2006, BMC Genomics, Vol. 7, p. 136. doi:10.1186/1471-2164-7-136.

12. *A metric for distributions with applications to image databases.* **Rubner, Yossi, Tomasi, Carlo and Guibas, Leonidas J.** Bombay, India : s.n., 1998. Proceedings of the 1998 IEEE International Conference on Computer Vision.

13. **Jensen, Paul A. and Bard, Jonathan F.** *Operation Research Models and Methods.* s.l. : John Wiley and Sons, 2003.

14. **(2012), R Development Core Team.** *R: A language and environment for statistical computing.* Vienna, Austria : R Foundation for Statistical Computing. URL http://www.R-project.org/. ISBN 3-900051-07-0.

15. *Significance analysis of microarrays applied to the ionizing radiation response.* **Tusher, Virginia Goss, Tibshirani, Robert and Chu, Gilbert.** 9, 2001, Proc Natl Acad Sci USA, Vol. 98, pp. 5116-5121. doi:10.1073/pnas.091062498.

16. *Bioconductor: open software development for computational biology and bioinformatics.* **Gentleman, Robert C., et al.** 2004, Genome Biology, Vol. 5, p. R80. URL http://genomebiology.com/2004/5/10/R80.

17. *Systematic and integrative analysis of large gene lists using David Bioinformatics Resources.* **Huang, D. W., Sherman, B. T. and Lempicki, R. A.** 1, 2009, Nature Protoc., Vol. 4, pp. 44-57.

18. *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* **Huang, D. W., Sherman, B. T. and Lempicki, R. A.** 1, 2009, Nucleic Acid Res., Vol. 37, pp. 1-13.

19. *Adjusting batch effect in microarray expression data using empirical Bayes methods.* **Johnson, W. Evan, Li, Cheng and Rabinovic, Ariel.** 1, 2007, Biostatistics, Vol. 8, pp. 118-127. doi:10.1093/biostatistics/kxj037.

20. **Leek, Jeffrey T., et al.** *sva: Surrogate Variable Analysis.* R package version 3.0.2.

21. *Gene-expression profiling of Waldenstrom macroglobulinemia reveals a phenotype more similar to chronic lymphocytic leukemia than multiple mieloma.* **Ching, Wee J., et al.** 2006, blood, Vol. 108, pp. 2755-2763. doi:10.1182/blood-2006-02-005488.

Section 9

APPENDIX

Table 10

CHIP ID	DIAGNOSIS	WM LABEL	BONE MARROW INFILTRATION [%]	SEX	AGE (at blood sample)	PATIENT ID
2001_02_22_WM03_CD19_U133_PLUS2	WM	1	90	F	79.5	WM03
2010_02_16_WM02_CD19	WM	1	90	F	53.9	WM02
2010_03_15_WM01_CD19_U133_PLUS2	WM	1	50	F	80.9	WM01
2010_03_25_WM04_CD19_U133_PLUS2	WM	2	30	F	77.2	WM04
2010_03_25_WM06_CD19_U133_PLUS2	WM	3	10	F	73	WM06
2010_03_31_WM07_CD19	WM	2	30	M	70.1	WM07
2010_03_31_WM08_CD19	WM	2	50	F	76.7	WM08
2010_04_08_WM09_CD19_PLUS2	WM	3	50	F	53.6	WM09
2010_05_07_WM13_19_U133_PLUS2	WM	3	10	F	65.8	WM13
2010_05_11_WM12_CD19_U133_PLUS2	IgM MGUS	5	2	M	65.6	MGUS19/WM12
2010_08_27_WM18_CD19_U133_PLUS2	WM	2	30	M	78.4	WM18
2010_08_27_WM19_CD19_U133_PLUS2	WM	2	70	F	65.7	WM19
2010_08_31_WM15_CD19_U133_PLUS2	WM	3	80	F	75.6	WM15
2010_09_03_WM21_CD19_U133_PLUS2	WM	3	20	F	73.7	WM21/MGUS11
2011_02_18_WM22_CD19_U133_PLUS2	IgM MGUS	5	9	F	76.5	MGUS12/WM22
2011_02_24_WM23_CD19_U133_PLUS2	WM	1	80	M	69.7	WM23
2011_03_01_MGUS02_CD19_U133_PLUS2	WM	2	90	F	80.5	WM33/MGUS02
2011_03_02_MGUS03_CD19_U133_PLUS2	WM	3	50	M	63.3	WM34/MGUS03
2011_05_17_WM24PV_CD19_U133_PLUS2	IgM MGUS	5	0	F	76.1	MGUS18/WM24PV
2011_05_20_WM25PV_CD19_U133_PLUS2	WM	3	45	M	64.3	WM25PV
2011_05_20_WM31_CD19_U133_PLUS2	WM	3	40	M	78	WM31/MGUS07
2011_06_09_WM26_CD19_U133_PLUS2	IgM MGUS	5	0	F	62.4	MGUS20/WM26
2011_06_09_WM27_CD19_U133_PLUS2	WM	2	30	M	86.7	WM27
2011_06_17_WM28_CD19_U133_PLUS2	WM	3	10	F	62.9	WM28
2011_09_20_WM35_CD19_U133_PLUS2	WM	3	70	F	61.9	WM35/MGUS14

2011_10_27_MGUS16_CD19_U133_PLUS2	IgM MGUS	5	0	M	72.3	MGUS16
2012_01_11_WM05_CD19_U133PLUS2	WM	3	10	F	71.3	WM05
2012_01_18_MGUS22_CD19_U133PLUS2	IgM MGUS	5	9	F	58.4	MGUS22/WM36
2012_04_11_MGUS24_CD19	IgM MGUS	5	0	M	75.9	MGUS24
2012_04_11_MGUS25_CD19	IgM MGUS	5	9	F	73.9	MGUS25
2012_04_11_MGUS28_CD19	IgM MGUS	5	0	M	75.4	MGUS28
2012_06_21_WM37_CD19	WM	3	30	M	77.8	WM37
2012_06_29_MGUS17_CD19_2	IgM MGUS	5	0	M	70.6	MGUS17
2012_06_29_MGUS21_CD19_2	IgM MGUS	5	9	F	80.5	MGUS21
2012_07_04_WM39_CD19	WM	3	30	M	69.7	WM39
2012_07_04_WM41_CD19	WM	2	50	M	55.4	WM41
2012_07_04_WM42_CD19	WM	3	50	M	74.3	WM42
2012_07_04_WM43_CD19	WM	3	10	F	70.5	WM43

Table 11

Annotation Cluster 1	**Enrichment Score: 1.9161**					
Category	Term	Count	%	PValue	Genes	FDR
GOTERM_MF_FAT	GO:0005539~glycosaminoglycan binding	6	4.0541	0.0038	LPL, APP, NOD2, SERPINE2, TGFBR3, PTCH1	0.0488
GOTERM_MF_FAT	GO:0001871~pattern binding	6	4.0541	0.0057	LPL, APP, NOD2, SERPINE2, TGFBR3, PTCH1	0.0721
GOTERM_MF_FAT	GO:0030247~polysaccharide binding	6	4.0541	0.0057	LPL, APP, NOD2, SERPINE2, TGFBR3, PTCH1	0.0721
GOTERM_MF_FAT	GO:0008201~heparin binding	5	3.3784	0.0071	LPL, APP, SERPINE2, TGFBR3, PTCH1	0.0897
GOTERM_MF_FAT	GO:0030246~carbohydrate binding	7	4.7297	0.0472	LPL, APP, NOD2, SERPINE2, CLEC2B, TGFBR3, PTCH1	0.4717
SP_PIR_KEYWORDS	heparin-binding	3	2.0270	0.0784	LPL, APP, SERPINE2	0.6499
Annotation Cluster 2	**Enrichment Score: 1.8333**					
Category	Term	Count	%	PValue	Genes	FDR
SP_PIR_KEYWORDS	actin-binding	10	6.7568	0.0001	PPP1R9A, LIMA1, SPTBN5, CNN3, DMD, CAPZA1, MARCKS, TPM2, PALLD, MYLK	0.0009
GOTERM_MF_FAT	GO:0003779~actin binding	10	6.7568	0.0007	PPP1R9A, LIMA1, SPTBN5, CNN3, DMD, CAPZA1, MARCKS, TPM2, PALLD, MYLK	0.0089
GOTERM_MF_FAT	GO:0008092~cytoskeletal protein binding	12	8.1081	0.0012	PPP1R9A, FRMD5, LIMA1, SPTBN5, CNN3, DMD, CAPZA1, MARCKS, TPM2, PALLD, FARP1, MYLK	0.0158
SP_PIR_KEYWORDS	cytoskeleton	12	8.1081	0.0058	PPP1R9A, LIMA1, SPTBN5, DMD, FRMD4B, CAPZA1, AKAP12, MARCKS, TPM2, PALLD, SEPT10, LATS2	0.0726
GOTERM_CC_FAT	GO:0005856~cytoskeleton	19	12.8378	0.0111	LIMA1, SPTBN5, CNN3, CAPZA1, PSD3, AKAP12, PALLD, TPM2, SEPT10, LATS2, FARP1, PPP1R9A, FRMD5, APP, HOMER3, DMD, FRMD4B, SYNJ2, MARCKS	0.1289
GOTERM_CC_FAT	GO:0015629~actin cytoskeleton	7	4.7297	0.0143	PPP1R9A, LIMA1, SPTBN5, CAPZA1, MARCKS, TPM2, PALLD	0.1638
GOTERM_CC_FAT	GO:0044430~cytoskeletal part	14	9.4595	0.0217	LIMA1, CNN3, SPTBN5, CAPZA1, PSD3, PALLD, TPM2, SEPT10, LATS2, APP, PPP1R9A, HOMER3, SYNJ2, MARCKS	0.2384
GOTERM_BP_FAT	GO:0030036~actin cytoskeleton organization	5	3.3784	0.0793	PPP1R9A, LIMA1, SPTBN5, CNN3, CAPZA1	0.7356
GOTERM_BP_FAT	GO:0007010~cytoskeleton organization	7	4.7297	0.0931	PPP1R9A, LIMA1, SPTBN5, CNN3, DMD, CAPZA1, PALLD	0.7924
GOTERM_BP_FAT	GO:0030029~actin filament-based process	5	3.3784	0.0950	PPP1R9A, LIMA1, SPTBN5, CNN3, CAPZA1	0.7993
GOTERM_CC_FAT	GO:0043228~non-membrane-bounded organelle	22	14.8649	0.3505	LIMA1, SPTBN5, CNN3, CAPZA1, PSD3, AKAP12, MYBL1, PALLD, TPM2, SEPT10, LATS2, FARP1, VDR, PPP1R9A, FRMD5, APP, HOMER3, DMD, FRMD4B, SYNJ2, MARCKS, MRO	0.9953
GOTERM_CC_FAT	GO:0043232~intracellular non-membrane-bounded organelle	22	14.8649	0.3505	LIMA1, SPTBN5, CNN3, CAPZA1, PSD3, AKAP12, MYBL1, PALLD, TPM2, SEPT10, LATS2, FARP1, VDR, PPP1R9A, FRMD5, APP, HOMER3, DMD, FRMD4B, SYNJ2, MARCKS, MRO	0.9953

Annotation Cluster 4	**Enrichment Score: 1.7209**					
Category	Term	Count	%	PValue	Genes	FDR
UP_SEQ_FEATURE	domain:Ig-like C2-type 3	6	4.0541	0.0019	VSIG10, SDK2, PALLD, MYLK, LAG3, FCRL3	0.0273
UP_SEQ_FEATURE	domain:Ig-like C2-type 4	5	3.3784	0.0022	VSIG10, SDK2, PALLD, MYLK, FCRL3	0.0318
INTERPRO	IPR013783:Immunoglobulin-like fold	12	8.1081	0.0026	VSIG10, IGHG1, LAIR1, IGHV1-69, IGHG3, SDK2, IGKV3D-15, LOC100133862, PALLD, IGHM, FCRL3, IGHV3-11, TRIM2, IGHV3-7, IGH@, IGHV4-31, LAG3, MYLK, LOC100126583	0.0349
INTERPRO	IPR007110:Immunoglobulin-like	11	7.4324	0.0040	VSIG10, IGHG1, IGHV1-69, IGHG3, SDK2, IGKV3D-15, LOC100133862, PALLD, IGHM, SLAMF1, FCRL3, IGHV3-11, IGHV3-7, IGH@, IGHV4-31, LAG3, MYLK, LOC100126583	0.0527
UP_SEQ_FEATURE	domain:Ig-like C2-type 5	4	2.7027	0.0072	SDK2, PALLD, MYLK, FCRL3	0.1016
INTERPRO	IPR013106:Immunoglobulin V-set	7	4.7297	0.0104	IGHG1, VSIG10, IGHV1-69, IGHG3, IGKV3D-15, LOC100133862, IGHM, FCRL3, IGHV3-11, IGHV3-7, IGH@, IGHV4-31, LAG3, LOC100126583	0.1322
UP_SEQ_FEATURE	domain:Ig-like C2-type 1	6	4.0541	0.0120	VSIG10, SDK2, PALLD, MYLK, LAG3, FCRL3	0.1637
UP_SEQ_FEATURE	domain:Ig-like C2-type 2	6	4.0541	0.0122	VSIG10, SDK2, PALLD, MYLK, LAG3, FCRL3	0.1669
INTERPRO	IPR013151:Immunoglobulin	6	4.0541	0.0171	VSIG10, IGHV3-11, IGHG1, IGHG3, IGHV3-7, SDK2, IGHV4-31, IGH@, IGHM, LAG3, FCRL3, LOC100126583	0.2071
SP_PIR_KEYWORDS	Immunoglobulin domain	9	6.0811	0.0194	IGHG1, VSIG10, LAIR1, IGHG3, SDK2, IGHM, PALLD, SLAMF1, FCRL3, IGHV3-11, IGHV3-7, IGH@, IGHV4-31, LAG3, MYLK	0.2230
INTERPRO	IPR003596:Immunoglobulin V-set, subgroup	4	2.7027	0.0200	IGHV3-11, IGHG1, IGHV1-69, IGHG3, IGHV3-7, IGKV3D-15, LOC100133862, IGHV4-31, IGH@, IGHM, LOC100126583	0.2387
UP_SEQ_FEATURE	domain:Ig-like C2-type 6	3	2.0270	0.0288	SDK2, MYLK, FCRL3	0.3520
SMART	SM00406:IGv	4	2.7027	0.0333	IGHV3-11, IGHG1, IGHV1-69, IGHG3, IGHV3-7, IGKV3D-15, LOC100133862, IGHV4-31, IGH@, IGHM, LOC100126583	0.3064
INTERPRO	IPR003598:Immunoglobulin subtype 2	5	3.3784	0.0655	VSIG10, SDK2, PALLD, MYLK, FCRL3	0.5991
INTERPRO	IPR003599:Immunoglobulin subtype	6	4.0541	0.0977	VSIG10, LAIR1, SDK2, MYLK, LAG3, FCRL3	0.7504
SMART	SM00408:IGc2	5	3.3784	0.1141	VSIG10, SDK2, PALLD, MYLK, FCRL3	0.7298
SMART	SM00409:IG	6	4.0541	0.1775	VSIG10, LAIR1, SDK2, MYLK, LAG3, FCRL3	0.8789
INTERPRO	IPR013098:Immunoglobulin I-set	3	2.0270	0.2712	SDK2, PALLD, MYLK	0.9860
Annotation Cluster 5	**Enrichment Score: 1.6320**					
Category	Term	Count	%	PValue	Genes	FDR
GOTERM_CC_FAT	GO:0005886~plasma membrane	46	31.0811	0.0001	IGHG1, IGHG3, LIMA1, GRIK1, IL15, IGHM, IL17RB, FCRL3, GNG8, APP, NOD2, FRMD5, HOMER3, SYNJ2, LAG3, PRKCA, LAIR1, PTPRN2, PSD3, PCDH9, MPP7, CCNL2, IGHV3-11, CCR6, CCR2, CYBRD1, AKAP7, IGH@, GPR183, CPM, AKAP12, CACNB2, NETO1, IGF1R, DMD, CLEC2B, CNR1, ZAP70, LPL, ADAM29, KL, SLAMF1, FZD7, IGHV3-7, PPP1R9A, GPR34, P2RX1, PON2, TGFBR3, IGHV4-31, PTCH1, FAM84B, PLEKHA1	0.0016

GOTERM_CC_FAT	GO:0044459~plasma membrane part	31	20.9459	0.0004	GPR183, LIMA1, GRIK1, CACNB2, IL15, IL17RB, GNG8, IGF1R, APP, HOMER3, DMD, CLEC2B, CNR1, ZAP70, SYNJ2, LAG3, ADAM29, KL, PTPRN2, PSD3, MPP7, SLAMF1, CCNL2, PPP1R9A, CCR6, GPR34, P2RX1, CCR2, CYBRD1, TGFBR3, AKAP7, PTCH1	0.0044
GOTERM_CC_FAT	GO:0005887~integral to plasma membrane	18	12.1622	0.0055	GPR183, ADAM29, GRIK1, KL, PTPRN2, CACNB2, IL15, CCNL2, IL17RB, IGF1R, APP, CCR6, GPR34, P2RX1, CNR1, CLEC2B, CCR2, TGFBR3, PTCH1	0.0667
GOTERM_CC_FAT	GO:0031226~intrinsic to plasma membrane	18	12.1622	0.0069	GPR183, ADAM29, GRIK1, KL, PTPRN2, CACNB2, IL15, CCNL2, IL17RB, IGF1R, APP, CCR6, GPR34, P2RX1, CNR1, CLEC2B, CCR2, TGFBR3, PTCH1	0.0826
SP_PIR_KEYWORDS	receptor	21	14.1892	0.0085	IGHG1, GPR183, IGHG3, GRIK1, IGHM, NETO1, FCRL3, IL17RB, VDR, IGF1R, CNR1, CRY1, LAIR1, PTPRN2, PKHD1L1, SLAMF1, FZD7, CCNL2, IGHV3-11, IGHV3-7, CCR6, GPR34, P2RX1, CCR2, TGFBR3, IGHV4-31, IGH@, PTCH1	0.1042
SP_PIR_KEYWORDS	glycoprotein	44	29.7297	0.0097	IGHG1, ARSD, IGHG3, GRIK1, IL15, IGHM, LASS6, IL17RB, FCRL3, APP, KCNK9, SERPINE2, DMKN, CNTNAP2, PDGFD, LAG3, VSIG10, LAIR1, PTPRN2, SDK2, PCDH9, PKHD1L1, CCNL2, IGHV3-11, UGT2B17, CCR6, CCR2, CYBRD1, IGH@, CPM, NETO1, IGF1R, CLEC2B, CNR1, FGL2, LPL, ADAM29, KLK2, KL, RNASE6, SLAMF1, LCN10, FZD7, IGHV3-7, GPR34, SNED1, P2RX1, PON2, TGFBR3, IGHV4-31, PTCH1	0.1174
SP_PIR_KEYWORDS	signal	35	23.6486	0.0115	IGHG1, CPM, IGHG3, ARSD, GRIK1, IL15, IGHM, FCRL3, NETO1, IL17RB, IGF1R, APP, SERPINE2, DMKN, CNTNAP2, FGL2, PDGFD, LAG3, VSIG10, LAIR1, LPL, ADAM29, KLK2, KL, PTPRN2, SDK2, RNASE6, PCDH9, PKHD1L1, LCN10, SLAMF1, FZD7, IGHV3-11, IGHV3-7, UGT2B17, SNED1, ITPRIPL2, TGFBR3, PON2, IGHV4-31, IGH@	0.1378
SP_PIR_KEYWORDS	cell membrane	26	17.5676	0.0117	IGHG1, GPR183, IGHG3, CPM, GRIK1, CACNB2, IGHM, FCRL3, IL17RB, NETO1, GNG8, HOMER3, DMD, CNR1, ZAP70, PRKCA, LPL, LAIR1, KL, PSD3, PCDH9, SLAMF1, CCNL2, IGHV3-11, IGHV3-7, CCR6, GPR34, CCR2, TGFBR3, AKAP7, IGHV4-31, IGH@, PLEKHA1	0.1409
UP_SEQ_FEATURE	signal peptide	35	23.6486	0.0126	IGHG1, CPM, IGHG3, ARSD, GRIK1, IL15, IGHM, FCRL3, NETO1, IL17RB, IGF1R, APP, SERPINE2, DMKN, CNTNAP2, FGL2, PDGFD, LAG3, VSIG10, LAIR1, LPL, ADAM29, KLK2, KL, PTPRN2, SDK2, RNASE6, PCDH9, PKHD1L1, LCN10, SLAMF1, FZD7, IGHV3-11, IGHV3-7, UGT2B17, SNED1, ITPRIPL2, TGFBR3, PON2, IGHV4-31, IGH@	0.1710
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	42	28.3784	0.0138	IGHG1, ARSD, IGHG3, GRIK1, IL15, IGHM, LASS6, IL17RB, FCRL3, APP, KCNK9, SERPINE2, CNTNAP2, PDGFD, LAG3, VSIG10, LAIR1, PTPRN2, SDK2, PCDH9, CCNL2, IGHV3-11, UGT2B17, CCR6, CCR2, CYBRD1, IGH@, CPM, NETO1, IGF1R, CLEC2B, CNR1, FGL2, LPL, ADAM29, KLK2, KL, RNASE6, SLAMF1, LCN10, FZD7, IGHV3-7, GPR34, SNED1, P2RX1, PON2, TGFBR3, IGHV4-31, PTCH1	0.1856
UP_SEQ_FEATURE	disulfide bond	30	20.2703	0.0265	IGHG1, GPR183, IGHG3, CPM, IL15, IGHM, FCRL3, NETO1, IGF1R, APP, CLEC2B, CNTNAP2, FGL2, PDGFD, LAG3, VSIG10, LAIR1, LPL, ADAM29, KLK2, SDK2, RNASE6, PALLD, LCN10, SLAMF1, FZD7, CCNL2, IGHV3-11, IGHV3-7, CCR6, GPR34, P2RX1, SNED1, CCR2, PON2, IGHV4-31, IGH@	0.3280
UP_SEQ_FEATURE	topological domain:Extracellular	29	19.5946	0.0288	GPR183, GRIK1, IL17RB, NETO1, FCRL3, IGF1R, APP, KCNK9, CLEC2B, CNR1, CNTNAP2, LAG3, VSIG10, LAIR1, ADAM29, KL, PTPRN2, SDK2, PCDH9, PKHD1L1, SLAMF1, FZD7, CCNL2, CCR6, GPR34, P2RX1, ITPRIPL2, CCR2, TGFBR3, PTCH1	0.3520
SP_PIR_KEYWORDS	membrane	56	37.8378	0.0360	IGHG1, IGHG3, GRIK1, VPS37B, IGHM, LASS6, IL17RB, FCRL3, GNG8, APP, FRMD5, KCNK9, HOMER3, SYNJ2, CNTNAP2, LAG3, VSIG10, PRKCA, LAIR1, PTPRN2, SDK2, PSD3, PCDH9, MPP7, PKHD1L1, CCNL2, FRY, IGHV3-11, TMEM133, UGT2B17, CCR6, CCR2, CYBRD1, AKAP7, IGH@, GPR183, CPM, CACNB2, NETO1, IGF1R, CNR1, CLEC2B, DMD, ZAP70, LAPTM4B, LPL, C1ORF162, ADAM29, KL, SLAMF1, FZD7, SH3BP4, IGHV3-7, GPR34, P2RX1, ITPRIPL2, MBOAT1, TGFBR3, PON2, PTCH1, MARCKS, IGHV4-31, PLEKHA1	0.3761
SP_PIR_KEYWORDS	disulfide bond	30	20.2703	0.0376	IGHG1, GPR183, IGHG3, CPM, IL15, IGHM, FCRL3, NETO1, IGF1R, APP, CLEC2B,	0.3895

					CNTNAP2, FGL2, PDGFD, LAG3, VSIG10, LAIR1, LPL, ADAM29, KLK2, SDK2, RNASE6, PALLD, LCN10, SLAMF1, FZD7, CCNL2, IGHV3-11, IGHV3-7, CCR6, GPR34, P2RX1, SNED1, CCR2, PON2, IGHV4-31, IGH@	
GOTERM_BP_FAT	GO:0007166~cell surface receptor linked signal transduction	20	13.5135	0.0669	GPR183, ADAM29, GRIK1, KL, KLF10, AKAP12, FZD7, CCNL2, GNG8, IGF1R, APP, CCR6, GPR34, HOMER3, CNR1, CCR2, ZAP70, TGFBR3, PTCH1, LAG3, PLEKHA1	0.6722
SP_PIR_KEYWORDS	transmembrane protein	9	6.0811	0.0884	IGF1R, GPR183, APP, CCR6, CNR1, PTPRN2, CCR2, TGFBR3, LAG3, CCNL2	0.6960
UP_SEQ_FEATURE	topological domain:Cytoplasmic	30	20.2703	0.1654	GPR183, GRIK1, LASS6, IL17RB, NETO1, FCRL3, IGF1R, APP, KCNK9, CLEC2B, CNR1, CNTNAP2, LAG3, VSIG10, LAIR1, ADAM29, KL, PTPRN2, SDK2, PCDH9, PKHD1L1, SLAMF1, FZD7, CCNL2, CCR6, GPR34, P2RX1, ITPRIPL2, CCR2, TGFBR3, PTCH1	0.9314
GOTERM_CC_FAT	GO:0031224~intrinsic to membrane	45	30.4054	0.2534	IGHG1, IGHG3, GRIK1, IL15, IGHM, LASS6, IL17RB, FCRL3, APP, FRMD5, KCNK9, CNTNAP2, LAG3, VSIG10, LAIR1, PTPRN2, SDK2, PCDH9, PKHD1L1, CCNL2, FRY, TMEM133, IGHV3-11, UGT2B17, CCR6, CCR2, CYBRD1, AKAP7, IGH@, GPR183, CPM, CACNB2, NETO1, IGF1R, CLEC2B, CNR1, LAPTM4B, LPL, C1ORF162, ADAM29, KL, SLAMF1, FZD7, IGHV3-7, GPR34, P2RX1, ITPRIPL2, MBOAT1, TGFBR3, IGHV4-31, MARCKS, PTCH1	0.9733
SP_PIR_KEYWORDS	transmembrane	39	26.3514	0.3372	IGHG1, GPR183, IGHG3, GRIK1, IGHM, LASS6, FCRL3, NETO1, IL17RB, IGF1R, FRMD5, APP, KCNK9, CNR1, CLEC2B, CNTNAP2, LAG3, VSIG10, LAPTM4B, LAIR1, C1ORF162, ADAM29, KL, PTPRN2, SDK2, PCDH9, PKHD1L1, SLAMF1, FZD7, CCNL2, FRY, IGHV3-11, TMEM133, IGHV3-7, UGT2B17, CCR6, GPR34, P2RX1, CCR2, ITPRIPL2, CYBRD1, MBOAT1, TGFBR3, PTCH1, IGHV4-31, IGH@	0.9950
UP_SEQ_FEATURE	transmembrane region	38	25.6757	0.3938	GPR183, GRIK1, LASS6, FCRL3, IL17RB, NETO1, IGF1R, FRMD5, APP, KCNK9, CLEC2B, CNR1, CNTNAP2, LAG3, VSIG10, LAPTM4B, LAIR1, C1ORF162, ADAM29, KL, PTPRN2, SDK2, PCDH9, PKHD1L1, SLAMF1, FZD7, CCNL2, FRY, TMEM133, UGT2B17, CCR6, GPR34, P2RX1, CCR2, ITPRIPL2, MBOAT1, CYBRD1, TGFBR3, PTCH1	0.9994
GOTERM_CC_FAT	GO:0016021~integral to membrane	41	27.7027	0.4521	IGHG1, GPR183, IGHG3, GRIK1, CACNB2, IL15, IGHM, LASS6, FCRL3, NETO1, IL17RB, IGF1R, FRMD5, APP, KCNK9, CNR1, CLEC2B, CNTNAP2, LAG3, VSIG10, LAPTM4B, LAIR1, C1ORF162, ADAM29, KL, PTPRN2, SDK2, PCDH9, PKHD1L1, SLAMF1, FZD7, CCNL2, FRY, IGHV3-11, TMEM133, IGHV3-7, UGT2B17, CCR6, GPR34, P2RX1, CCR2, ITPRIPL2, CYBRD1, MBOAT1, TGFBR3, PTCH1, IGHV4-31, IGH@	0.9994
Annotation Cluster 7	**Enrichment Score: 1.3862**					
Category	Term	Count	%	PValue	Genes	FDR
GOTERM_CC_FAT	GO:0043005~neuron projection	9	6.0811	0.0037	IGF1R, APP, PPP1R9A, KIAA1598, GPR34, CNN3, GRIK1, SYNJ2, APBB2	0.0444
GOTERM_CC_FAT	GO:0042995~cell projection	10	6.7568	0.0702	IGF1R, APP, PPP1R9A, KIAA1598, GPR34, CNN3, GRIK1, CYBRD1, SYNJ2, APBB2	0.5948
GOTERM_CC_FAT	GO:0044463~cell projection part	5	3.3784	0.0938	APP, PPP1R9A, CNN3, CYBRD1, SYNJ2	0.7054
GOTERM_CC_FAT	GO:0030425~dendrite	4	2.7027	0.1185	PPP1R9A, GPR34, CNN3, GRIK1	0.7908