Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Unsupervised analysis of leukemia and normal hematopoiesis by joint clustering of gene expression data

Liviu Badea

Bioinformatics Group, National Institute for Research in Informatics
8-10 Averescu Blvd., Bucharest, Romania

***Abstract*—Leukemia is a very heterogeneous cancer of the hematopoietic system. Since its main cause consists of genomic defects in the hematopoietic stem or progenitor cells and given the high complexity of the hematopoietic system, it may seem an important task to investigate the transcriptomic similarities and differences between leukemia subtypes and hematopoietic cells (stem cells, progenitors and differentiated cells).**

In this paper, we integrate the largest publicly available gene expression datasets of leukemia and normal hematopoiesis with the aim of uncovering the main gene modules involved in normal hematopoiesis as well as in the various leukemia subtypes.

Using a joint consensus clustering algorithm, we have been able to relate the major leukemia types to their putative cells of origin in an unsupervised manner. While the normal hematopoietic cell modules are also active in leukemias of the corresponding cell type, our approach has determined leukemia-specific modules comprising genes with a known involvement in leukemogenesis.

The expression modules uncovered implicate an unusually large number of transcription factors. This speaks against very simple models of normal hematopoiesis and leukemogenesis that involve just a handful of critical TFs, arguing for the interplay of complex transcription factor networks, in line with the findings of the FANTOM consortium for leukemia and Novershtern et al. for normal hematopoiesis.

***Index Terms*—joint clustering, leukemia, hematopoiesis.**

I. Introduction

Leukemia is a very heterogeneous cancer of the hematopoietic system, which involves complex genomic changes in hematopoietic stem cells, leading to abnormalities in various hematopoietic cell populations. Its current classification includes acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myelogeneous leukemia (AML) and chronic myelogeneous leukemia (CML), with a myriad of subtypes and other rarer types. Since its main cause consists of genomic defects in the hematopoietic stem or progenitor cells and given the high complexity of the hematopoietic system, it may seem an important task to investigate the transcriptomic similarities and differences between leukemia subtypes and hematopoietic cells (stem cells, progenitors and differentiated cells).

Given the unique experimental accessibility of the various cell compartments, including the stem cells, leukemias are probably one of the best suited cancers for a genomic experimental investigation of the cancer stem cell hypothesis. A related hypothesis suggesting that cancer involves developmental programs gone awry can also be tested, as long as detailed transcriptomic data about normal hematopoietic differentiation is available.

In this paper, we integrate the largest publicly available gene expression datasets of leukemia and normal hematopoiesis with the aim of uncovering the main gene modules involved in normal hematopoiesis as well as in the various leukemia subtypes. The main assumption is that leukemia reuses "normal" gene modules in inappropriate ways – finding these modules and associating them to leukemia subtypes is of great importance for developing a more detailed genomic subclassification of leukemia. This is needed because the heterogeneity of most subtypes is not entirely accounted for by the current classification, which is based mainly on the cell types and the more frequent genomic changes. For example, in the case of AML, the majority of cases are classified as 'AML with normal karyotype and other abnormalities', while the cases with more frequent well defined translocations represent just a minority. Thus, although the current classification can be accurately predicted from genomic data based on *supervised* machine learning methods [9], it seems of importance to be able to characterize the main disease subtypes as well as the associated gene modules in an *unsupervised*, or *semi-supervised* manner.

One of the simplest and most effective clustering methods for microarray data is based on *Nonnegative Matrix Factorization* (*NMF*) [14], [5], [11], which tends to produce sparse and domain-interpretable decompositions using a very simple computational framework. The nonnegativity constraints imposed by NMF distinguish it from other dimensional reduction methods such as Principal Component Analysis, which tend to produce more "holistic" decompositions that are much harder to relate to real biological sub-processes.

While a large number of gene expression studies employing matrix factorization in general and NMF in particular have been put forward (with [5], [11] among the first), only a much smaller number of studies were able to deal with *simultaneous factorizations* of several relations (or matrices):

- Alter et al. [1] have employed the generalized SVD (GSVD) algorithm for comparing two cell cycle datasets (sharing the sample timepoints),
- Lee et al. [13] also used GSVD, but this time for comparing array CGH copy number profiles from patient-matched normal and tumor samples,
- Ponnapalli et al. [20] generalized GSVD to a Higher-

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

Order GSVD (which allows an arbitrary number of matrices sharing a given dimension) and applied it again to cell cycle data,

- Badea [3] introduced the siNMF algorithm which simultaneously factorizes a pair of gene expression matrices with matching genes. [2] further generalizes siNMF to an arbitrary multirelational setting involving an arbitrary number of entities linked by binary relations.

Unfortunately, none of these approaches is without problems. Like their original SVD counterpart, GSVD-based approaches tend to produce holistic decompositions which are hard to interpret or to relate to precise biological networks. On the other hand, NMF-based approaches are prone to *instability*, especially in multi-relational domains, where different runs of the algorithm (with different initializations) tend to produce distinct results (clusters).

We have developed a *multirelational consensus clustering* method [submitted] that is able to deal with the inherent instability of multirelational clustering. In this paper, we apply it to the unsupervised joint subclassification of leukemias and normal hematopoiesis. For this, we have combined the largest publicly available transcriptomic dataset for leukemia, the MILE study [9] with the largest gene expression study of normal hematopoiesis [18]. For a relatively small number of target clusters (namely 15), the algorithm was able to recover in an unsupervised manner the main types of leukemia and normal hematopoietic cells, as well as to link major leukemia types to their putative cells of origin.

II. Methods

A. The datasets

We have combined the gene expression data of the MILE leukemia study [9] with the transcriptomic data for normal hematopoiesis of [18].

The *MILE study* has measured gene expression profiles from the bone marrow (1556 samples) and peripheral blood (540 samples) of 2096 patients: 750 ALL cases, 542 AML, 448 CLL, 76 CML, 206 MDS (myelodysplastic syndrome), 74 healthy persons. Each sample was assigned to one of 17 more detailed leukemia subtyes or to 'normal'. Unfortunately, no follow-up information was available for this very large cohort of patients (T. Haferlach, personal communication). The raw Affymetrix U133 Plus 2.0 CEL files were downloaded from GEO (datset GSE13159) and reprocessed using the RMA algorithm implemented in Affymetrix Power Tools (APT).

The study of Novershtern et al. [18] has profiled the transcriptomes of 38 distinct types of purified hematopoietic cells (211 replicates in total) on a slightly different microarray platform (Affymetrix U133A). We have downloaded the raw CEL files from GEO (dataset GSE24759) and RMA-normalized them also using APT.

Since virtually all U133A probesets are also present on the U133 Plus 2.0 chip, we have retained only the common probesets on the two platforms (22268 probesets). We then filtered the probesets retaining only those with a significant expression (mean of the log_2-values $> log_2(100)$ and standard deviation of log_2 values > 0.8). This resulted in 7417 probesets.

Besides the gene expression matrices of the leukemia (X_L) and respectively hematopoiesis dataset (X_H), we employed the given subtype information, Y_L for leukemia and Y_H for hematopoiesis. $Y_L(s_L, t_L)$ is 1 if leukemia sample s_L is of subtype t_L and 0 otherwise. (Similarly for Y_H.) There are 18 leukemia subtypes and 38 subtypes of normal hematopoietic cells.

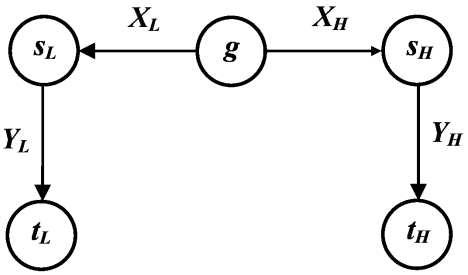

Fig. 1. The joint analysis of leukemia (L) and hematopoiesis (H). g : genes, s_L, s_H : samples, t_L, t_H : subtypes, X_L, X_H : gene expression matrices, Y_L, Y_H : subtype matrices

B. Multirelational clustering with Nonnegative Matrix Factorization

We briefly review the framework fo multirelational learning using nonnegative decompositions [2].

A *multirelational domain* involves a set of *entity types* $\{\mathcal{E}^{(n)}\}_n$ as well as a set of numerical relations $\{R^{(mn)}\}_{mn}$ between these entity types. An *entity type* $\mathcal{E}^{(n)}$ is a set of N_n related entities (such as genes, samples or disease subtypes). In our setting, the nonnegative real-valued relation matrices $R_{ij}^{(mn)}$ are weighted by means of *weight matrices* $W_{ij}^{(mn)}$, which allow us to represent unknown relation entries (i, j) (by setting $W_{ij}^{(mn)} = 0$), as well as to balance relations with widely disparate value ranges.

Figure 1 presents the multirelational domain of interest in this paper. The gene expression matrices for leukemia (X_L) and hematopoiesis (X_H) are viewed as numerical relations between genes (g) and samples (s_L, s_H). Note that since genes represent a shared entity type, the corresponding gene clusters will be common to the factorizations of X_L and X_H.

More precisely, a rank-N_c *multirelational nonnegative decomposition* (MNMF) of a multi-relational structure $\langle\{\mathcal{E}^{(n)}\}_n, \{R^{(mn)}\}_{mn}, W^{(mn)}\}_{mn}\}\rangle$ is an assignment of a nonnegative factor matrix $E^{(n)}$ of size $N_n \times N_c$ to each entity type $\mathcal{E}^{(n)}$, such that all relations $R^{(mn)}$ are approximated by the product of the corresponding entity type matrices

$$R^{(mn)} \approx E^{(m)} \cdot E^{(n)T}. \tag{1}$$

Formally, we minimize the following weighted squared error function

$$f = \frac{1}{2} \sum_{s,d} \sum_{i,j} W_{ij}^{(sd)} \Big(R_{ij}^{(sd)} - \sum_{c=1}^{N_c} E_{ic}^{(s)} \cdot E_{jc}^{(d)} \Big)^2 \tag{2}$$

subject to nonnegativity constraints for the entity matrices $E^{(n)} \geq 0$.

A simple algorithm solving the optimization problem (2) was developed by generalizing the method employed by Lee and Seung for standard NMF [14]. The algorithm randomly

initializes the entity matrices $E^{(n)}$ and then iteratively applies the following multiplicative update rules until convergence:

$$E^{(n)} \leftarrow E^{(n)} \circ \frac{N^{(n)}}{P^{(n)}}, \text{ with} \tag{3}$$

$$\begin{aligned} P^{(n)} &= \sum_{(s,n)\in\mathcal{R}} \left[W^{(sn)} \circ \left(E^{(s)} \cdot E^{(n)T}\right)\right]^T \cdot E^{(s)} \\ &\quad + \sum_{(n,d)\in\mathcal{R}} \left[W^{(nd)} \circ \left(E^{(n)} \cdot E^{(d)T}\right)\right] \cdot E^{(d)} \end{aligned} \tag{4}$$

$$\begin{aligned} N^{(n)} &= \sum_{(s,n)\in\mathcal{R}} \left[W^{(sn)} \circ R^{(sn)}\right]^T \cdot E^{(s)} \\ &\quad + \sum_{(n,d)\in\mathcal{R}} \left[W^{(nd)} \circ R^{(nd)}\right] \cdot E^{(d)} \end{aligned} \tag{5}$$

where '$\circ$' and '$-$' represent elementwise (Hadamard) multiplication and respectively division of matrices, while $(m,n) \in \mathcal{R}$ denotes the existence of a relation between entity types $\mathcal{E}^{(m)}$ and $\mathcal{E}^{(n)}$.

Elsewhere [submitted] we prove that the error function (2) is nonincreasing under the multiplicative update rules (3).

C. Multirelational consensus clustering

Clustering gene expression data is affected by the small sample sizes compared to the numbers of variables, which leads to clustering *instability*. *Consensus clustering* aims at obtaining clusters that are more stable across different clustering runs. However, developing a consensus clustering algorithm for multi-relational decompositions is non-trivial. Existing consensus clustering approaches [16] construct a consensus matrix of items, which records for each item pair the frequency of their co-occurrence in the same cluster during a number of different clustering runs. Unfortunately, this simple idea only works for one-way clustering and not for the biclusters (two-way clusters) produced by (multi-relational) matrix factorizations. To deal with this problem, in [4] we have used a *Positive Tensor Factorization* [25] for clustering the biclusters obtained in a number of different factorization runs. Furthermore, we generalized this approach to the multi-relational setting [submitted].

Briefly, we start with a number N_r of different runs of the multirelational MNMF algorithm, which is assumed to have produced N_r individual factorizations $\{E_r^{(n)}\}_{\substack{n=1,\ldots,N_e \\ r=1,\ldots,N_r}}$ (index n refers to the entity type, while r refers to the run). $E_r^{(n)}$ are entity matrices whose entries $E_{icr}^{(n)}$ denote the membership of entity i (having entity type n) to cluster c of run r.

A *consensus clustering* corresponds to

- a set of *consensus entity matrices* $e_{ik}^{(n)}$ (with i an entity and $k \in \{1,\ldots,N_c\}$ an index referring to a specific *consensus cluster*), together with
- a *cluster correspondence array* α_{crk} (which shows how the individual clusters c from run r are recomposed from consensus clusters k)

such that the biclusters obtained in the different runs can be recovered from the following Positive Tensor Factorization:

$$E_{icr}^{(s)} \cdot E_{jcr}^{(d)} \approx \sum_{k=1}^{N_c} \alpha_{crk} e_{ik}^{(s)} e_{jk}^{(d)}. \tag{6}$$

More formally, (6) is rewritten as a minimization problem for the following error function:

$$F\left(\alpha, \{e^{(n)}\}_n\right) = \frac{1}{2} \sum_{\substack{(s,d)\in\mathcal{R} \\ c,r,i,j}} \left(E_{i(cr)}^{(s)} E_{j(cr)}^{(d)} - \sum_{k=1}^{N_c} \alpha_{(cr)k} e_{ik}^{(s)} e_{jk}^{(d)}\right)^2. \tag{7}$$

Note that in (7) we have grouped the (cr) indices in α and E in order to deal with matrices rather than 3-dimensional arrays.

The objective function (7) above aims at minimizing the Euclidean distance between the bicluster c from run r (given by $\left(E_{i(cr)}^{(s)} E_{j(cr)}^{(d)}\right)_{ij}$) and the cluster reconstructed from the consensus biclusters $\left(e_{ik}^{(s)} e_{jk}^{(d)}\right)_{ij}$ by means of the cluster correspondence matrix $\alpha_{(cr)k}$.

The *consensus clustering algorithm* runs MNMF N_r times, randomly initializes $\{e^{(n)}\}_n$ and α, then iteratively applies the following update rules

$$e^{(n)} \leftarrow e^{(n)} \circ \frac{E^{(n)} \cdot \left(\alpha \circ \sum_{(d,n)\text{or}(n,d)\in\mathcal{R}} E^{(d)T} \cdot e^{(d)}\right)}{e^{(n)} \cdot \left((\alpha^T \cdot \alpha) \circ \sum_{(d,n)\text{or}(n,d)\in\mathcal{R}} e^{(d)T} \cdot e^{(d)}\right)}$$

$$\alpha \leftarrow \alpha \circ \frac{\sum_{(s,d)\in\mathcal{R}} \left(E^{(s)T} \cdot e^{(s)}\right)\left(E^{(d)T} \cdot e^{(d)}\right)}{\alpha \cdot \sum_{(s,d)\in\mathcal{R}} \left(e^{(s)T} \cdot e^{(s)}\right)\left(e^{(d)T} \cdot e^{(d)}\right)}$$

until convergence. Subsequently, α is normalized such that $\sum_{c,r} \alpha_{(cr)k} = N_r$. Finally, the consensus clusters $\{e^{(n)}\}_n$ are used as initialization for a final MNMF run.

Note that the consensus clusters need not necessarily be highly recurring clusters across the different runs. They could form a "base" set of clusters out of which all the clusters could be reconstructed by means of linear combinations. This allows learning of frequently occurring *subclusters*, thereby alleviating the need for very large numbers of runs.

We applied our consensus clustering algorithm to the relational structure from Figure 1. We used relation weights to equalize the Euclidean norms of the relations and then reduced the weights of the subtype relations by $1/100$ to avoid any significant bias of the known subtype information on the inferred clusters.

Next, we determined the number of clusters N_c based on a series of MNMF runs with progressively larger N_c, ranging from 2 to 50 (see Figure 2). To avoid overfitting, we performed a similar set of runs on the randomized entity matrices and compared the decrease of the error with N_c in the two cases. An N_c was chosen such that the error decrease on the real data was significantly larger than that on the randomized data [11]. In the following, we have chosen $N_c = 15$ clusters.

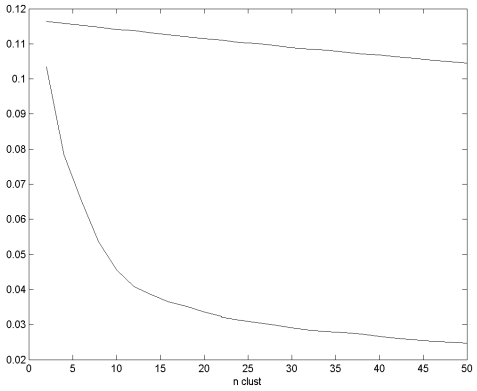

Fig. 2. The decomposition error as a function of the number of clusters for real and randomized data

D. Cluster annotation

Clusters obtained with our consensus clustering algorithm were analysed in detail using several annotation tools. We used the DAVID functional annotation tool (v6.7 online at david.abcc.ncifcrf.gov) to obtain the most significant annotations for the top 100 genes of each cluster. We also used the FANTOM4 EdgeExpressDB database for gene regulation in acute myeloid leukemia (online at fantom.gsc.riken.jp/4/edgeexpress) to construct putative gene regulation networks corresponding to the top 50 and respectively 100 genes of each cluster. EdgeExpressDB networks for all gene modules can be consulted as *supplementary information* online at ai.ici.ro/bibe2012. For the normalized gene cluster matrix $E^{(1)}$, a relatively strict significance threshold was employed: $\frac{2}{\sqrt{N_1}}$.

III. Results

As intended, the multirelational consensus clustering algorithm tends to infer sample-specific *gene modules* (biclusters) rather than obtain a simple unidimensional clustering of the samples. These modules may be involved both in disease and in normal cells, although some modules are specific to leukemia and others to normal hematopoietic cells.

Figures 3 and 4 show the sample clusters for leukemia and respectively normal hematopoiesis. (Rows correspond to samples, while columns represent clusters.) Note that certain clusters overlap, indicating the activation of several gene modules in the corresponding samples. At the 15 cluster-level of granularity, the clustering easily recognizes the main leukemia classes: T-ALL, B precursor ALL, CLL, AML, while CML and especially MDS and normals are clustered closer together. Other gene modules/clusters are more specific to normal hematopoiesis, with major distinctions between mature B cells, T cells and cells of the myeloid lineage respectively.

In the following, we present a more detailed analysis of the clusters obtained.

Cluster 1 represents a gene module activated mainly in *B precursor ALL* (c-ALL/pre-B-ALL, pro-B-ALL with t(11q23)/MLL, ALL with t(12;21), ALL with hyperdiploid karyotype and ALL with t(1;19)), which involves less differentiated B cells. Although significantly active mainly in leukemia, this gene module is also weakly active in the corresponding normal hematopoietic cells, namely early B cells, pro B cells or even hematopoietic stem cells (both CD133+CD34dim and CD38-CD34+). The functional annotation of the module with DAVID revealed genes involved in the *immune response* (p-value $7.5 \cdot 10^{-9}$): HLA-DQB1, CIITA, POU2AF1, HLA-DRB1, RAG1, PAX5, IGHM, HLA-DMA, CD74, LAT2, LILRA2, CD79B, DEFA1, HLA-DPA1, HLA-DPB1, CD24, BLNK, HLA-DRA, as well as in *lymphocyte activation* ($p = 3.3 \cdot 10^{-5}$): LAT2, RAG1, SOX4, BANK1, CD24, HLA-DMA, CD74, BLNK.

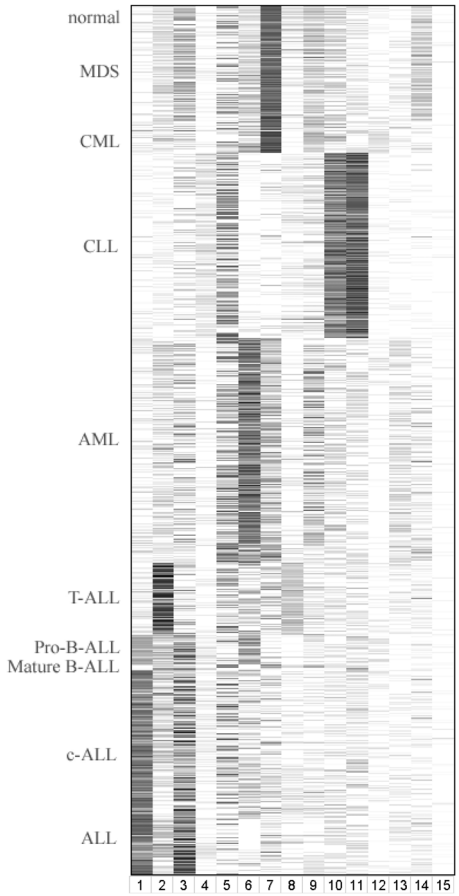

Fig. 3. The leukemia sample clusters

Comparing cluster 1 with gene module 4 (specific, as we shall see below to normal mature B cells) revealed a significant overexpression of SOCS2 and of the transcription factors SOX4 and PAX5. Note that SOX4 was recently shown to *cause* acute leukemia if overexpressed in mouse hematopoietic stem cells [23]. PAX5 encodes the B-cell lineage specific activator that is expressed at early stages of B-cell differentiation. PAX5 is a critical factor in B-ALL development and aberrant PAX5 expression especially at early stages may lead to leukemic transformation [8]. The EdgeExpressDB network associated to the top 50 genes of cluster 1 is depicted in Figure 5.

Gene modules 10 and 11 are mainly active in chronic lymphocytic leukemia (CLL). However, while module 10 seems to be driven mainly by TCF4, module 11 is probably controlled by a more complex network including several transcription factors, such as ID3, PAX5, KLF4, KLF6, KLF9, LEF1, JUN, etc. (see the corresponding EdgeExpressDB networks from the

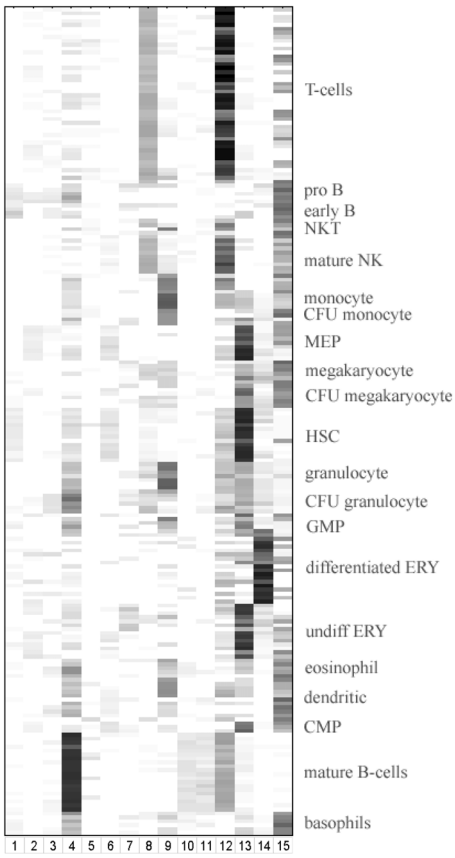

Fig. 4. The hematopoiesis sample clusters

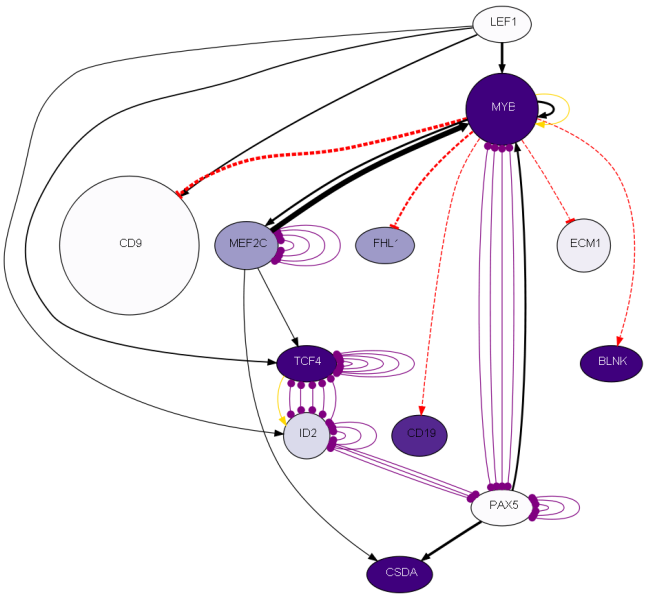

Fig. 5. EdgeExpressDB network for cluster 1 (B precursor ALL)

supplementary information online). PAX5 has been previously observed to be expressed in all B-CLL and pre-B ALL [17], precisely confirming our findings (see the involvement of PAX5 in the B precursor ALL module 1, as well as its central role in the EdgeExpressDB networks for clusters 1 and 11). Although primarily active in CLL, these two gene modules (10 and 11) are also weakly active in mature B-cells, as expected.

As opposed to the gene modules discussed above, *module 4* is primarily activated in normal mature B-cells, but is also weakly involved in CLL, as expected. According to EdgeExpressDB, it is mainly driven by IRF8, MEF2C and TCF4 (also see supplementary information online). Significant annotations for this cluster included *lymphocyte activation* ($p = 7.5 \cdot 10^{-9}$): EGR1, PTPRC, BCL11A, MS4A1, SMAD3, BANK1, CD24, TPD52, HLA-DMA, CD74, BLNK, and *immune response* ($p = 2.3 \cdot 10^{-19}$).

Gene module 6 is dominant mainly in AML cases, but is also weakly active in hematopoietic stem cells (HSC CD133+CD34dim and CD38-CD34+), megakaryocyte/erythroid progeniors (MEP) and common myeloid progenitors (CMP). Its closest normal counterpart is *gene module 13*, which is primarily expressed in HSC, MEP, CMP, as well as in the least differentiated erythroid progenitors (CD34+CD71+GlyA- and CD34-CD71+GlyA-), with weaker activation in the more differentiated progenitors. Apparently, gene module 13 encodes a stem cell/progenitor-specific expression program (L2L annotations: hsc_hsc and progenitors_adult, $p = 7.7 \cdot 10^{-9}$) involving PBX1, MYB, GATA2, ETS2, TAL1, etc.

Compared to module 13, *module 6* overexpresses key transcription factors like SOX4, HOXA10, CITED2, JUN, FOS, etc. Although the normal function of SOX4 in hematopoietic stem cells (HSCs) is not known, its overexpression in mouse HSCs has been shown to cause myeloid leukemia [23]. It is also known that deregulation of HOXA10 initiates AML [21], as HOXA10 is a critical regulator of HSCs and erythroid/megakaryocyte development [15]. CITED2 is an essential regulator of adult HSCs [12]. It has also been previously demonstrated that JUN shows an elevated expression in AML [22], as does FOS [24].

The most significant cluster 6 gene (with the largest coefficient) is FLT3, a receptor tyrosine kinase that regulates hematopoiesis. Mutations that result in the constitutive activation of this receptor result in acute myeloid leukemia and acute lymphoblastic leukemia. Even in the absence of mutations, overexpression of FLT3 has been shown to associate with a poor prognosis for overall survival [10].

Significant annotations for module 6 included *bone marrow stem progenitor* ($p = 5 \cdot 10^{-11}$): CEBPA, HIST2H2AA3, GNA15, LMO2, LGALS1, ANXA1, SPINK2, RPL36, IGF2BP2, TAGLN2, FAM46A, AZU1, HHEX, FOS, TARP, HIST2H2BE, HOXA10, CAT, PRKACB, RPL10A, CFD, SRGN, ATP8B4, and *bone marrow acute myelogenous leukemia* ($p = 5.1 \cdot 10^{-9}$).

In contrast, cluster 13 showed, as expected, a significant annotation for *normal bone marrow* ($p = 5.9 \cdot 10^{-6}$).

It is remarkable that the highest level stem cell in the hematopoietic lineage (CD133+CD34dim) is primarily involved in *acute* leukemias (B precursor ALL in module 1 and respectively AML in module 6).

Gene module 2 covers the T-ALL cases, while its "normal" counterpart, *module 8* is mainly active in normal T-cells and certain NK cells, with weaker activation in T-ALL. Comparing

the T-ALL expression program (module 2) with the normal T-cell expression module 8, we noticed increased expression of SOX4, MYB, JUN and TOP2A in T-ALL. Interestingly, the MYB trascription factor oncogene is tandemly duplicated in T-ALL [19] and may represent a novel therapeutic target [6]. Note the prominent role of MYB in B-recursor ALL (module 1), T-ALL (module 2) and AML (module 6) (cf. EdgeExpressDB networks in the supplementary information).

Gene module 14 covers normal differentiated erythroid cells (annotation *erythrocyte differentiation* $p = 10^{-6}$) and is only weakly active in MDS cases. Among the specific genes for this module are transcription factors like KLF1 (a known hematopoietic TF of adult erythroid genes), NFE2, as well as hemoglobin genes (HBA1, HBG1, HBB, etc.) and glycophorins (GYPA, GYPB – major sialoglycoproteins of the human erythrocyte membrane), etc.

Finally, *gene module 3* contains many genes involved in *cell division* ($p = 3.3 \cdot 10^{-8}$), e.g. CCNB1, CDK1, KIF11, MAD2L1, CETN3, PAFAH1B1, NDC80, ANAPC10, SMC1A, RACGAP1, SMC2, SMC4, and is active in most leukemia samples.

IV. Discussion

We have developed a method for inferring the main gene modules involved in leukemia and normal hematopoiesis. While the normal hematopoietic cell modules are also active in leukemias of the corresponding cell type, our approach has determined *leukemia-specific modules* involving genes with a known involvement in leukemogenesis, such as FLT3, HOXA10, SOX4, PAX5, MYB, etc. It is noteworthy that the algorithm has been able to relate the major leukemia types to their putative cells of origin in an *unsupervised* manner.

A careful analysis of the clusters obtained as well as a brief inspection of Figure 2 shows that a clustering with $N_c = 15$ clusters only reveals the main gross gene$\times$sample modules in the data. Such a coarse decomposition was crucial for validating our approach[1], but may need to be refined for obtaining finer-grained gene modules and disease subtypes. However, the validation of such finer grained decompositions would involve extensive experimental efforts and is therefore beyond the scope of this paper.

The expression modules uncovered involve an unusually large number of transcription factors. More precisely, using a relatively strict significance threshold for the normalized gene cluster matrix[2], we obtained 273 transcription factors significantly involved in the $N_c = 15$ clusters, a much larger number than normally expected. This speaks against very simple models of normal hematopoiesis and leukemogenesis that involve just a handful of critical TFs, arguing for the interplay of complex transcription factor networks, in line with the findings of the FANTOM consortium for leukemia [7] and [18] for normal hematopoiesis.

Finally, we believe that the generality of our multi-relational clustering algorithm will find numerous applications in the analysis of various high throughput data, given the urgent need to incorporate in a mathematically coherent way as much relational information about the problem as possible. After all, biological function is to a large extent relational.

[1] as the clusters obtained were coarse enough to enable a direct comparison with the known leukemia types

[2] We have normalized the columns of the gene cluster matrix to unit Euclidean norm.

Acknowledgment

This research was partially supported by the project PN-II-ID-PCE-2011-3-0198. I am grateful to Andrei Halanay, Daniel Coriu and Jardan Dumitru for discussions.

References

[1] O. Alter, et al. Generalized Singular Value Decomposition for Comparative Analysis of Genome-Scale Expression Datasets of Two Different Organisms, PNAS 100 (6), 33516, 2003.
[2] L. Badea. Multi-relational factorizations for cancer subclassification, Proc. ICACTE-2010, V1-248-252, 2010.
[3] L. Badea. Extracting Gene Expression Profiles Common to Colon and Pancreatic Adenocarcinoma Using Simultaneous Nonnegative Matrix Factorization. Proc. Pacific Symp. on Biocomputing 2008, 267-278.
[4] L. Badea. Clustering and Metaclustering with Nonnegative Matrix Decompositions. Proc. ECML-2005:10-22, 2005.
[5] J.P. Brunet, et al. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. 101: 4164-4169, 2004.
[6] Cools J. Identification and characterization of novel oncogenes in chronic eosinophilic leukemia and T-cell acute lymphoblastic leukemia. Verh K Acad Geneeskd Belg. 2010;72(1-2):55-70.
[7] FANTOM Consortium. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet. 2009 May;41(5):553-62.
[8] Firtina S, et al. Evaluation of PAX5 gene in the early stages of leukemic B cells in the childhood B cell acute lymphoblastic leukemia. Leuk Res. 2012 Jan;36(1):87-92.
[9] Haferlach T, et al. Global approach to the diagnosis of leukemia using gene expression profiling. Blood. 2005 Aug 15;106(4):1189-98.
[10] Kiyoi H, et al. Clinical significance of FLT3 in leukemia. Int J Hematol. 2005 Aug;82(2):85-92.
[11] Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13(7), 1706-18, 2003.
[12] Kranc KR, et al. Cited2 is an essential regulator of adult hematopoietic stem cells. Cell Stem Cell. 2009 Dec 4;5(6):659-65.
[13] C.H. Lee, et al. GSVD Comparison of Patient-Matched Normal and Tumor aCGH Profiles Reveals Global Copy-Number Alterations Predicting Glioblastoma Multiforme Survival, PLoS One 7(1):e30098, 2012.
[14] Lee DD and Seung HS. Algorithms for non-negative matrix factorization. in *NIPS*, pp. 556–562, 2000.
[15] Magnusson M, et al. HOXA10 is a critical regulator for hematopoietic stem cells and erythroid/megakaryocyte development. Blood. 2007 May 1;109(9):3687-96.
[16] S. Monti, et al. Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data, Journal of Machine Learning, 52(1-2), 2003.
[17] Nishii K, et al. Expression of B cell-associated transcription factors in B-cell precursor acute lymphoblastic leukemia cells:Int J Hematol. 2000 Jun;71(4):372-8.
[18] Novershtern N, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144(2):296-309, 2011.
[19] O'Neil J, Look AT. Mechanisms of transcription factor deregulation in lymphoid cell transformation. Oncogene. 2007 Oct 15;26(47):6838-49.
[20] S.P. Ponnapalli, et al. A Higher-Order Generalized Singular Value Decomposition for Comparison of Global mRNA Expression from Multiple Organisms, PLoS One 6 (12), e28072, December 2011.
[21] Quere R, et al. High levels of the adhesion molecule CD44 on leukemic cells generate acute myeloid leukemia relapse after withdrawal of the initial transforming event. Leukemia. 2011 Mar;25(3):515-26.
[22] Rangatia J, et al. Elevated c-Jun expression in acute myeloid leukemias inhibits C/EBPalpha DNA binding via leucine zipper domain interaction. Oncogene. 2003 Jul 24;22(30):4760-4.
[23] Richter K, et al. Global gene expression analyses of hematopoietic stem cell-like cell lines with inducible Lhx2 expression. BMC Genomics. 2006 Apr 6;7:75.
[24] P.B. Staber, et al. Common alterations in gene expression and increased proliferation in recurrent acute myeloid leukemia, Oncogene 23(4), 894-904, 2004.
[25] Welling M., Weber M. Positive tensor factorization. Pattern Recognition Letters 22(12): 1255-1261 (2001).