A prize-collecting Steiner tree application for signature selection to stratify diffuse large B-cell lymphoma subtypes

Authors: Murodzhon Akhmedov[1,2,4,*], Luca Galbusera[5,*], Roberto Montemanni[2], Francesco Bertoni[1,3] and Ivo Kwee[1,2,4,6]

Affiliations: [1]Università della Svizzera italiana, Institute of Oncology Research, Bellinzona, CH; [2]Dalle Molle Institute for Artificial Intelligence Research, SUPSI, Manno, CH; [3]Oncology Institute of Southern Switzerland, Bellinzona, CH; [4]Swiss Institute of Bioinformatics, Lausanne, CH; [5]University of Basel, Biozentrum, Basel, CH;[6]BigOmics Analytics, Bellinzona,CH; [*]Equal contributor

Current address: BigOmics Analytics, Bellinzona, CH (I.K, M.A.)

Correspondence: ivo.kwee@ior.usi.ch

ABSTRACT

Background: With the explosion of high-throughput data available in biology, the bottleneck is shifted to effective data interpretation. By taking advantage of the available data, it is possible to identify the biomarkers and signatures to distinguish subtypes of a specific cancer in the context of clinical trials. This requires sophisticated methods to retrieve the information out of the data, and various algorithms have been recently devised.

Results: Here, we applied the prize-collecting Steiner tree (PCST) approach to obtain a gene expression signature for the classification of diffuse large B-cell lymphoma (DLBCL). The PCST is a network-based approach to capture new insights about genomic data by incorporating an interaction network landscape. Moreover, we adopted the ElasticNet incorporating PCA as a classification method. We used seven public gene expression profiling datasets (three for training, and four for testing) available in the literature, and obtained 10 genes as signature. We tested these genes by employing ElasticNet, and compared the performance with the DAC algorithm as current golden standard. The performance of the PCST signature with ElasticNet outperformed the DAC in distinguishing the subtypes. In addition, the gene expression signature was able to accurately stratify DLBCL patients on survival data.

Conclusions: We developed a network-based optimization technique that performs unbiased signature selection by integrating genomic data with biological networks. Our classifier trained with the obtained signature outperformed the state-of-the-art method in subtype distinction and survival data stratification in DLBCL. The proposed method is a general approach that can be applied on other classification problems.

Key words: DLBCL, PCST, signature, high-throughput data

BACKGROUND

Diffuse large B-cell lymphoma (DLBCL) is the most common form of human lymphoma. Based on gene expression profiling studies it has been proposed that this clinically heterogeneous entity can be divided in at least two main classes, the germinal center B-cell (GCB) and the activated B-cell like (ABC), subtypes, depending on their phenotypic homologies with their supposed cell of origin (COO) [21, 22]. These subgroups have shown differences in clinical course and response to standard or experimental therapies, and they should allow the development of tailored therapies [21, 22]. The definition of the COO is now recommended in the clinical practice [23] although it still a matter of research on the identification of the best markers to be used and on the choice of the best technology to measure them [21, 24, 25].

Modern advances in microarrays and genome sequencing instruments have now provided the researchers with a huge amount of data from which it is possible to extract useful gene expression signatures to discriminate the two DLBCL subtypes. Nonetheless, the analysis of gene expression patterns is characterized by a large number of covariates, the genes analyzed, and a small number of observations, patients with known ABC or GCB DLBCL, and, thus, data mining must be accomplished with careful statistical techniques to identify, among all the analyzed genes, a signature including only the elements most useful for the class prediction.

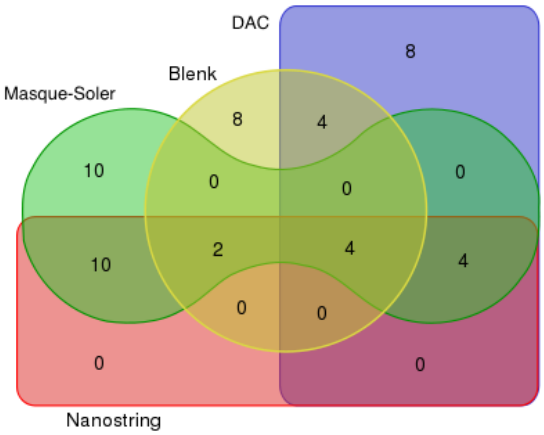

Figure 1: A Venn diagram demonstrating the overlap of four published signatures (DAC [13], Blenk [15], Masque-Soler [27], and Nanostring [33]). There is a considerable discordance in the gene lists. The reader is referred to the appendix for more information about the signatures analyzed in this paper.

There are many proposed signatures able to identify the two sub-classes of lymphoma [13, 14, 15, 26, 27, 28, 29, 32], but the lists of genes are discordant, as can be seen by the little overlap in three of them, shown in Fig.~1, highly dependent on the specific study and sometimes determining discordant COO definition of the same sample. We believe that one of the reasons for such heterogeneity lies in the preliminary filtering process, when a subset of all the analyzed genes is arbitrarily chosen to reduce the degrees of freedom and allow the application of a more robust statistical analysis. Here, we propose a method based on a Prize-collecting Steiner Tree (PCST), able to extract the important genes, searching in the space of all the available transcripts, thus removing the initial arbitrary filtering step. Basically the PCST approach enables the analysis of genomics data on a network level leading to the identification of a subnetwork that potentially can act as gene expression signature distinguishing the two DLBCL subtypes. We constructed gene interaction networks using gene expression data without applying any gene filtering as a whole. The nodes and edges of the resulting gene interaction network are given a score as the differential expression and cost, respectively. The cost of an edge is reversely proportional to the mutual information of expression values of corresponding genes. The cost of an edge is higher if the mutual information is smaller for that specific edge. After scoring the network, the PCST is applied to find a neighborhood of genes in the network that are differently expressed and connected to each other through the highly mutual informative (∼reliable) edges. The set of differently expressed genes that are collected with high mutual information among themselves would potentially stratify the DLBCL patients.

Materials and Methods

Datasets

We used publicly available datasets downloaded from Gene Expression Omnibus[1] (GEO) repository and reported in Table~1. Three datasets were used as training set and four for validation purposes.

To prepare the datasets for subsequent analysis, we transformed the probe names to gene names, removing non-annotated probes; if more than one probe was associated to the same gene, only the probe

[1] http://www.ncbi.nlm.nih.gov/geo/

Table 1: Publicly available datasets. They have been conducted with several different experimental conditions.

DATASET	PLATFORM	SAMPLES	USE	TYPE	AUTHOR
GSE22470	Hgu133A	176	Train	FF	Salaverria I.[43]
GSE19246	Hgu133Plus2	144	Train	Both	Williams P.M.[42]
GSE31312	Hgu133Plus2	451	Train	FFPE	Frei E.[41]
GSE4475	Hgu133A	96	Test	FF	Hummel M.[40]
GSE10172	Hgu133A	10	Test	FF	Klapper W.[39]
GSE23501	Hgu133Plus2	57	Test	FF	Shaknovich R.[38]
GSE10846	Hgu133Plus2	350	Test	FFPE	Lenz G.[37]

with the highest standard deviation was retained. In case of missing data, genes lacking measurement in more than 50% of the samples were removed, while for the other genes, missing values were imputed using a nearest neighbor average.

Table 2: DLBCL datasets used for training the method.

Dataset	Total samples	ABC samples	GCB samples
GSE19246	144	63	81
GSE22470	215	71	144
GSE31312	451	214	237

Networks

We generated interaction networks for the training data sets with up to 21,049 nodes and 1,423,267 edges (see Table~3) by using the multiplicative model of ARACNE algorithm, a tool for the reconstruction of gene interaction networks [16]. We generated two interaction networks for each training data with parameters $eps = 0.01$ and $eps = 0.05$ that fed into ARACNE (six interaction networks in total for three training data). In these networks, every edge represented the interaction between two genes. ARACNE scores the edges using mutual information of expression values of genes. Each node was labeled with prize $p_v = |E_{|ABC|} - E_{|GCB|}|$, where $E_{|ABC|}$ and $E_{|GCB|}$ are the mean value of gene expression of ABC and GCB cancer patients for corresponding gene, respectively. Each edge was labeled with a cost that was reverse proportional to the mutual information obtained by ARACNE. The cost was smaller if the mutual information was higher for that edge, and vice-versa otherwise.

Table 3: The sizes of network instances generated by from the training datasets.

	$eps = 0.001$		$eps = 0.005$	
Dataset	Nodes	Edges	Nodes	Edges
GSE19246	20639	82282	20639	154469
GSE22470	12994	266085	12994	375097
GSE31312	20639	1055055	20639	1423267

PCST - the network optimization method

The PCST is a well-known problem in graph theory and there are different variants of the problem in the literature [1, 2, 31]. Given an undirected graph $G = (V, E)$, where the vertices are labeled with prizes $p_v \geq 0$ and the edges are labeled with cost $c_e > 0$, the goal is to identify a subnetwork $G' = (V', E')$ with the tree structure. The target is to minimize the sum of the prizes of vertices out of V', and to minimize the total cost of the edges in E'. This is equivalent to the minimization of the following objective function [17]:

$$f(G') = \text{minimize} \sum_{e \in E'} c_e + \sum_{v \notin V'} p_v \tag{1}$$

The PCST has a wide range of applications, mainly in the design of utility networks such as fiber optic and district heating networks. Recently, some PCST applications in biological networks have been published [3, 10, 11, 12]. The biological graphs such as genetic interaction networks can be scored by using experimental data including gene expression profiling, mutation, or copy number data. In these networks, every vertex is a gene, and every edge between two vertices represents the genetic or protein interaction between the two vertices. Each vertex in the network is given a prize, which can be the differential expression between two subtypes or number of mutation for that gene [6] . Also, each edge in the network is given a cost, which corresponds to pairwise mutual information of expression values between two genes. After having scored all vertices and edges in the network, the PCST is employed to detect a relevant subnetwork (tree). Biologically, the subnetwork obtained by the PCST has an important meaning [3] , in which it corresponds to a portion of interaction network where many genes are closely related in terms of their functions, and potentially belong to the same biological pathways [6]. Furthermore, the resulting subnetwork contains many genes that are highly correlated and differently expressed between the sub- types in the experimental data. Therefore, the set of genes included into PCST solution tree can represent gene expression signatures to distinguish disease subtypes.

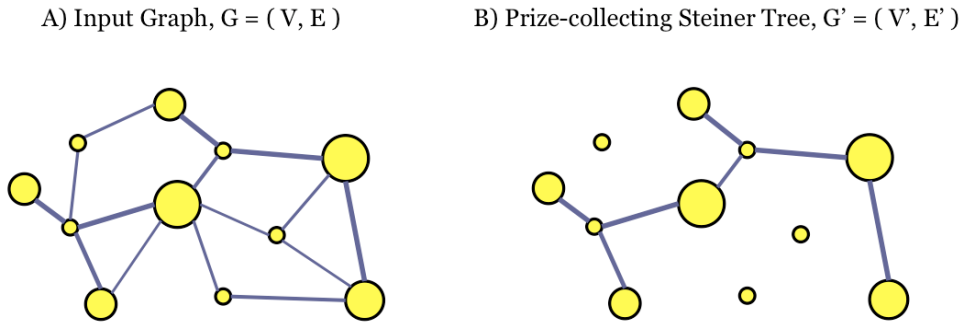

Figure 2: The Prize-collecting Steiner Tree problem instance and solution. A) An input graph instance. The node sizes and arc widths are proportional to the prizes and mutual information, respectively. B) The Prize-collecting Steiner Tree solution for input graph. The PCST collected set of genes that have higher differential expression and connected through the least cost (∼highly mutual informative) edges.

The illustration of the PCST on input graph and the resulting output tree are demonstrated in Fig.~2. The vertex sizes and edge widths are proportional to prize of nodes and mutual information of edges, respectively.

In the application of PCST in genomics, it is possible to have very large networks to analyze. Nevertheless, PCST has NP-hard characteristics which makes it computationally challenging for existing exact algorithms to find solutions in reasonable execution time. This yields a need for fast matheuristic algorithms to compute the PCST solution tree in large networks for inferring biologically relevant subnetworks. Considering challenges, previously we have developed a matheuristic approach [1, 2] which efficiently scales up on large biological instances. The matheuristic was composed of the preprocessing procedures, a heuristic clustering algorithm and an exact PCST solver from the literature [31]. Here, we employ a prior approach developed in [1] to compute the PCST solution tree for biological networks.

Signature

We employed the PCST method [1] on six interaction networks generated by ARACNE [16] to obtain tree solutions. The PCST tree solution includes itself the group of genes that are differently expressed in ABC versus GCB subtype, and highly correlated with each other at the same time. Therefore, these genes were assumed promising as gene expression signature to stratify ABC and GCB subtypes.

Training the classification methods

We use Elastic Net as a classification method to distinguish DLBCL subtypes. The classification method is trained by using the signature obtained by PCST (see Results), and we examined the performance on test data to objectively evaluate it.

Penalized linear regression and mixture of experts

The idea is to look for a simple linear classifier performing a logistic regression. The Steiner tree allowed us to reduce the number of covariates from tens of thousands of genes to ten, however among them there could still be correlated genes or genes not relevant for the classification, which can lead to a sub-optimal estimation of the regression parameters. To select relevant genes several techniques of variable selection and penalized linear regressions have been devised to identify the important covariates. We used ElasticNet regression [20] that combines a continuous variable selection, while keeping under control the correlations among the covariates. The penalization introduced by ElasticNet can be seen in a Bayesian framework as the introduction of a prior distribution on the values of the coefficients which models our prior assumption that the value of the coefficients must be bounded (this keeps under control any correlation among the regressors) and that the solution could be sparse (variable selection). This penalization is an attempt to combine the well-known Lasso and Ridge methods, trying to overcome their individual limitations.

Moreover, we adopted a meta-analytic approach in which each training dataset is used to build an independent classifier. Each classifier performs a binary classification through a logistic regression and for every sample it casts a vote between 0 and 1 that can be interpreted as the probability for the sample to be of sub-type ABC. We tried two different ways to combine the votes of the single classifiers together. Either we used the mean value or the median; we found that the mean leads to best results (see Results and Supplementary Material).

Another question is how to present the training data to the classifiers. In fact, PCA might allow to reduce the dimensionality of the datasets and improve the prediction power of the ElasticNet model. As it can be seen in Fig.~3, if we consider the three training datasets restricted to the genes present in the signature found, the first three components are enough to explain more than 60% of the total variance, so we chose to consider only the first three components.

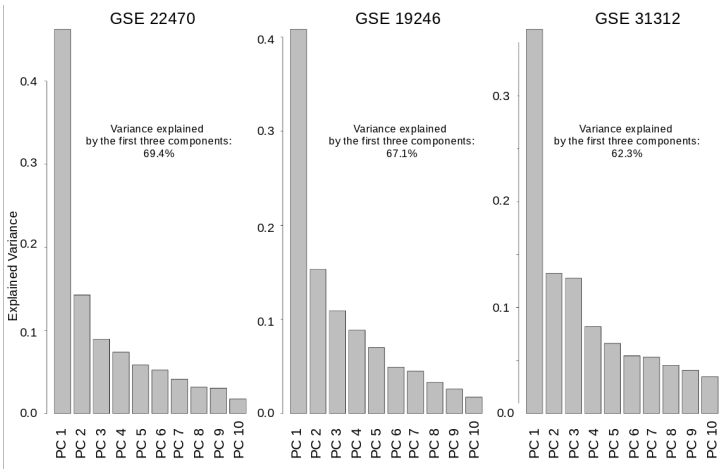

Figure 3: The PCA of the three training datasets shows that the first three components are enough to explain more than 60% of the total variance.

We tried to feed the classifier using either the real data or the first two principal components and we found that the best results are obtained using both representations (See Results and Supplementary Material for a more detailed discussion).

In the end the classification algorithm works as follows:

1. For each training dataset, train an ElasticNet classifier;
2. For each training dataset, train second ElasticNet classifier based on the first three principal components;
3. Let each classifier cast a vote for each sample in the test dataset;
4. Merge the votes together using the mean value.

Comparison with previous studies

We compared our gene expression signature and classifier with previously proposed ones. The comparison was on two phases: comparison of the classification accuracy and patient survival stratification ability.

To test whether our signature and classification algorithm can improve the present classification system, we tested it against the DLBCL Automatic Classifier (DAC) [13] one of the "state-of-the-art" classifiers for the DLBCL classification.

Since ABC and GCB DLBCL groups have different clinical outcome we compared the stratification using our classes and using the classes assigned by the original study, for two of the test datasets with available survival information.

Results and Discussion

The set of genes are selected as signature by implementing the PCST. We tested the performance of our signature and classifier on test datasets for stratification of patient subtypes and survival data. The results of the proposed method were very similar if we shuffled the datasets among the test and training sets.

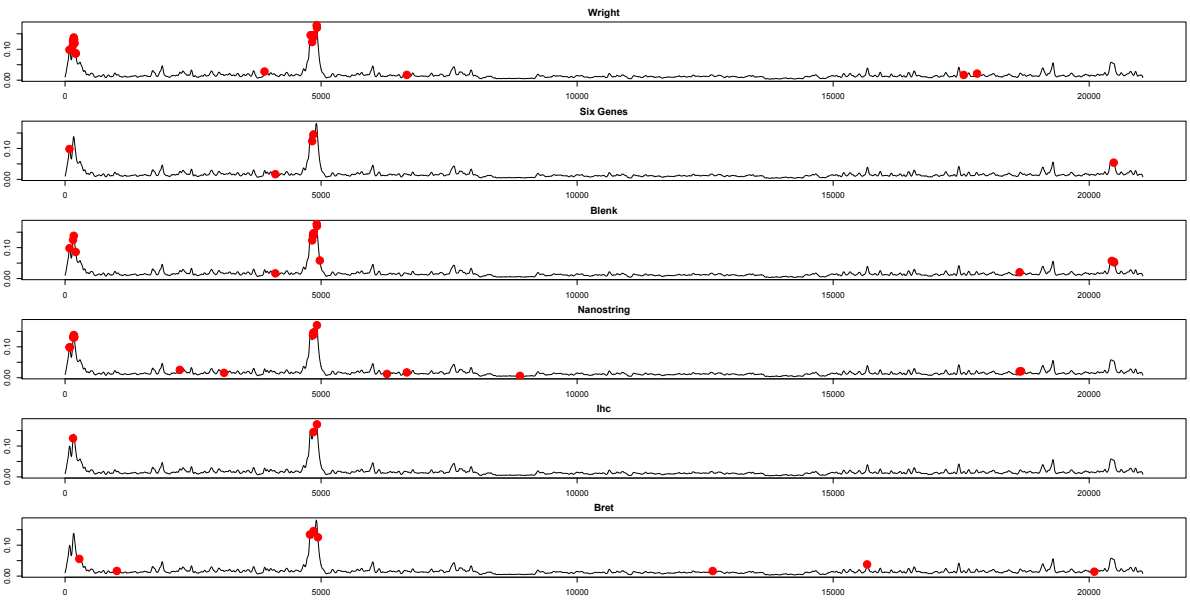

Figure 4: Fold change of all the genes ordered according to a clustering based on the correlation of their expression. The Wright signature is the golden standard today, while the IHC is the golden standard for diagnostic purposes.

Comparison to previous signatures

As discussed in the introduction, in the literature there are several signatures proposed to classify the ABC and the GCB subtypes of DLBCL. These signatures hardly overlap, but we will show that the proposed genes are highly correlated. Using the R library *nclust* [18, 19], we performed a hierarchical clustering of all the datasets we have; using one-minus-Pearson's correlation as the distance matrix. We then computed the fold-changes of each gene in the two classes of lymphoma and we ordered them according to the cluster tree, obtaining a signal exhibiting some peaks, smoothed with a Gaussian kernel, corresponding to genes highly correlated and differentially expressed in the two DBLCL classes. In Fig.~4 we show how the genes from the published signatures are distributed among the peaks of the ordered fold-changes. Different signatures, even if they do not overlap for specific genes, have several genes belonging to the same peaks, i.e. very correlated (see the appendix for information about the signatures). This strongly confirms the hypothesis that the differences in the signatures are due only to genes belonging to the same functional group, representing the same biologic function. Interestingly, the genes used for the immunoistochemical analysis fall exactly on the top of the peaks. We can see that our approach using the clustering to reorder the genes offers a unified vision of the different signatures.

The PCST signature

Applying the PCST method discussed above, we obtained output subnetworks for six interaction networks. These outputs were combined into a single component by taking the union of subnetworks as it is displayed in Fig.~5. The resulting network had 174 genes and 240 interactions. To get more robust signature set, we selected intersecting genes among all the subnetwork as signature, which composed of ten genes including MYBL1, PHF16, SSBP2, SPINK2, RAB7L1, STAG3, C13ORF18, BATF, MME, LMO2. Fig.~6 shows how our signature overlapped the main two peaks in the fold change of the cluster-ordered genes.

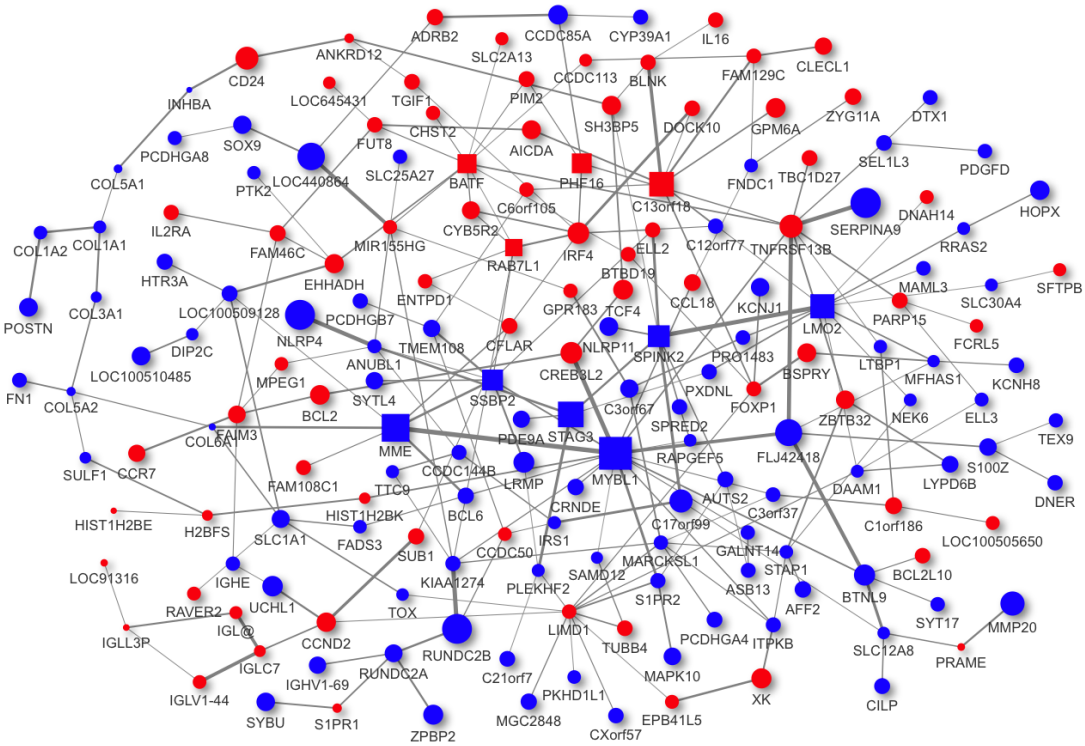

Figure 5: The union of the subnetworks obtained by the PCST method. The size of nodes and width of edges are proportional to the differential expression of genes and mutual information between the genes, respectively. Red colored genes are over-expressed in ABC, and blue colored genes are over-expressed in GCB subtype. The square shaped nodes are selected as signature to distinguish the subtypes.

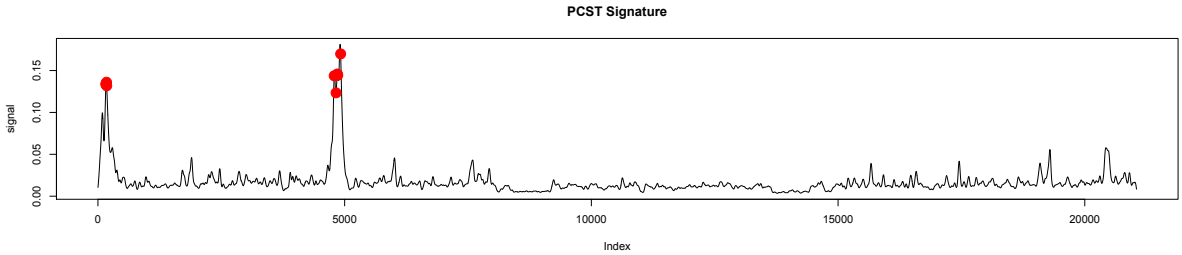

Figure 6: Our signature corresponds to genes belonging to the two main peaks of the cluster-ordered genes. These main peaks are also shared by the other signatures found in the literature (cfr. Fig. 4).

Furthermore, the expression levels of genes belonging to the gene expression signatute were compared between ABC and GCB subtypes. The genes were differentially expressed between the subtypes in the training datasets. The box plots of the expressions are provided in Fig.~ 7, 8 and 9.

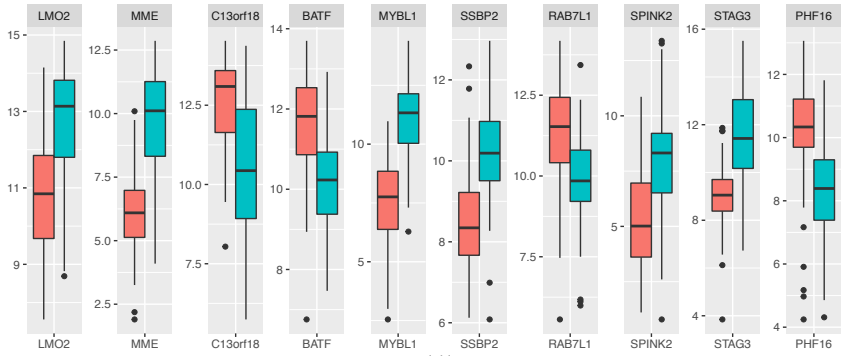

Figure 7: The mean values of ABC (orange) versus GCB (turquoise) subtype expressions for the PCST signature genes in GSE19246 dataset.

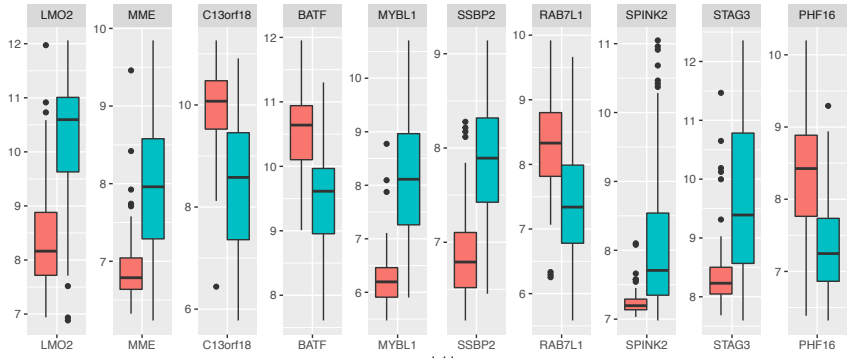

Figure 8: The mean values of ABC (orange) versus GCB (turquoise) subtype expressions for the PCST signature genes in GSE22470 dataset.

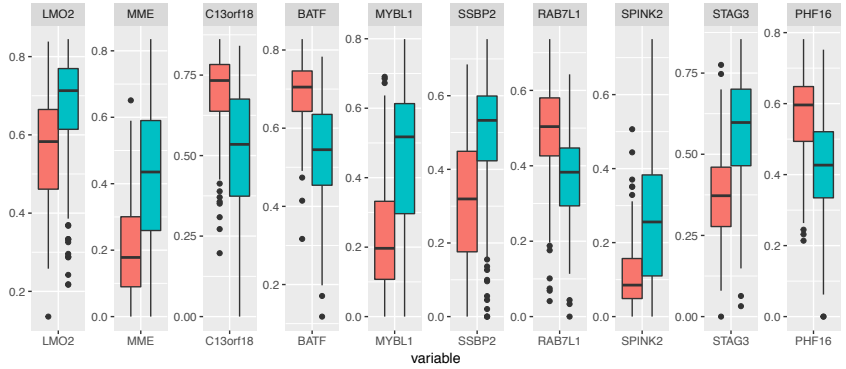

Figure 9: The mean values of ABC (orange) versus GCB (turquoise) subtype expressions for the PCST signature genes in GSE31312 dataset.

Accuracy comparison with previous models

The accuracy performance of our method was compared with the DAC classifier, which is a "gold standard" approach in the literature [13]. The accuracy was computed by dividing the number of correctly predicted patients by the number of total patients. In addition to ABC and GCB subtypes, the DAC classifier also makes the type-III class predictions. In order to make a fair comparison platform, we disregarded the type-III class predictions of DAC while computing the accuracy, or reassigned the type-III to ABC or GCB group with the highest prediction probability.

The comparison results on four test sets are reported in Table~4: our algorithm was robust against different datasets. Our method provided comparable results on the first two datasets with the accuracy of the DAC [13] classifier, and it significantly outperformed the DAC on the last two datasets. We used the procedure described in Methods, but in the Supplementary material section we showed the results of our classifier trained with different settings.

Table 4: Comparison of the DAC [13] algorithm with our method. For each dataset, we report the number of ABC and GCB samples, and the classification accuracy using DAC [13] or our algorithm.

Dataset	ABC	GCB	DAC*	DAC**	PCST+ENet
GSE4475	58	120	96.00	**97.00**	96.88
GSE10172	3	28	**90.00**	89.00	**90.00**
GSE10846	167	183	85.00	90.00	**96.57**
GSE23501	20	40	88.00	94.00	**100.00**

* type-III class prediction is disregarded
** type-III class prediction is reassigned to ABC or GCB that has the highest prediction probability

Survival analysis with the PCST signature

To provide a stronger evidence of relevance of PCST signature in clinical trials, we tested it in the analysis of survival data. The signature was used to classify DLBCL patients employing the survival data of test sets. Two out of four datasets came along with the survival data in GEO repository. The survival analyses are displayed in Fig.~10, where our method provided better p-values (2.56e-9 versus 1.14e-9 for GSE10846, 0.00296 versus 0.0000984 for GSE4475) in the stratification of DLBCL patients compared to previous approaches.

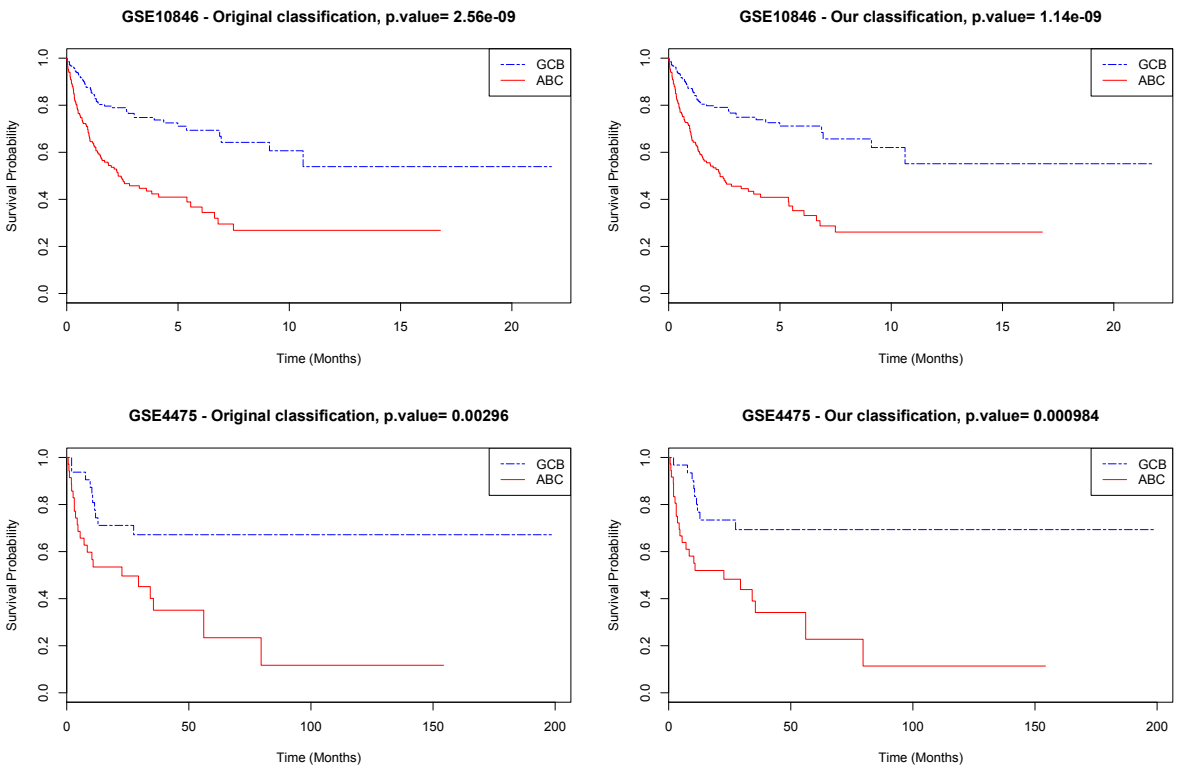

Figure 10: Survival analysis for two datasets: top row GSE4475, bottom row GSE10846. The left column shows the survival using the classes provided by the authors of the study, the right column shows the results using our classification.

Conclusion

Genomics technique is a promising technology to allow personalized medicine. However, the path to make them readily available for the everyday clinical practice is still far to be completed.

One main challenge is to identify the set of important classifier genes among the thousands of genes measured. Here, we proposed a method based on the well-known Steiner Tree algorithm. The genes analyzed by the specific platform are mapped onto the vertices of a genetic network where the vertex prize is given by the differential expression of the gene between the two subtypes of disease (hence it is a measure of how good the gene is able to discriminate the two subtypes) and the edges are given a cost corresponding to the mutual information between two genes connected by the vertex. The aim of the PCST is to find a relevant subnetwork with minimal cost; this subnetwork is comprised of genes with high discriminative power and mutual information. Since the problem is NP-hard, we employed a metaheuristic previously developed in our group [1, 2].

% In practice, we applied our method to a set of DLBCL samples in order to classify them in the two subtypes ABC and GCB, obtaining a signature of 10 genes. % The next step is implementing a classifier using the found signature. We chose to use a linear classifier with the so-called Elasticnet prior which is able to efficiently cope with correlated variables and possibly null true coefficients. We applied Elasticnet both on the real data and on the first two principal components, using 3 different training datasets. Following a metanalytic approach, we created a different classifier for the 6 training datasets (3 different sets, each considered in the real space and as the first two principal components). We then merged the votes from each classifier considering the mean score.

The gene expression signature obtained by the proposed method using the expression data was tested against the sate-of-the-art approaches in the literature. The PCST + ElasticNet+PCA classification

accuracies on test sets were equivalent or better than the gold-standard method. Besides, using the survival data, our method lead to a better stratification of DLBCL patients.

Finally, the proposed feature selection method is robust for noise in data due to the batch effects and other environmental factors. Furthermore, it is very generic that can be easily applied on other types of cancer data for signature selection and subtype stratification.

Supplementary

Publicly available signatures

In the literature, various signatures have been proposed in order to classify the DLBCL. In Fig.~4 we showed how the genes from these signatures overlap with the peaks of the hierarchical clustering. The signatures used are

1. Wright: One of the first signatures [34].
2. Six Genes [35]: Minimal signature composed of only six genes
3. Blenk [15]: Signature of 18 genes
4. Nanostring [33]: Signature used to classify FFPE tissues using the Nanostring technology.
5. IHC [14]: Signature for IHC classification
6. Bret [36]: 12 genes signature

Using different parameters to train the classifier

In the main test we showed the classification accuracy using a mixture of experts composed on models trained on the real data and on the first three principal components. The final vote is given by the mean of the single experts.

However, one can wonder if using both real data and the first three principal components is not redundant, or if the median is better than the mean since it is less dependent on outliers. We present in Table~5 the results using either the mean or the median on classifiers trained only with real data, only with the three first principal components or both. As it can be seen, the mean values of classifiers trained both on real data and on the principal components allows to achieve the best performance. We also report the results obtained with the DAC [13] classifier.

Table 5: Comparison of the mixture of experts using ElasticNet. The different mixtures depend on how the training data are feed into the classifiers and how the final vote is obtained as a combination of the votes of the single classifiers.

			DAC [13]		PCST + {ElasticNet + PCA}					
					Mean			Median		
Dataset	ABC	GCB	*	**	Both	PCA	Real	Both	PCA	Real
GSE4475	58	120	96.00	**97.00**	96.88	93.75	94.79	95.83	96.88	94.79
GSE10172	3	28	**90.00**	89.00	**90.00**	80.00	**90.00**	80.00	80.00	**90.00**
GSE10846	167	183	85.00	90.00	**96.57**	94.57	**96.57**	96.00	95.14	95.71
GSE23501	20	40	88.00	94.00	**100.00**	92.98	94.74	96.49	92.98	94.74

* type-III class prediction is disregarded

** type-III class prediction is reassigned to ABC or GCB that has the highest prediction probability

Abbreviations

PCST: Prize-collecting Steiner Tree; GEO: Gene Expression Omnibus; DLBCL: Diffuse large B-cell lymphoma; ABC: Activated B-cell like; GCB: Germinal center B-cell; COO: Cell of origin; DAC: DLBCL Automatic Classifier

Acknowledgements

Not applicable

Funding

This work was supported by the Swiss National Science Foundation throughout the project 205321-147138/1 and the GELO Foundation.

Availability of data and materials

All the gene expression data used during the current study is publicly available on Gene Expression Omnibus. The software for developed signature selection and classification is available from the corresponding author on request.

Author's contributions

MA and LG designed the experiments, performed all analysis and wrote the paper. MA, IK and RM developed the PCST algorithm. IK and FB conceived the experiments and provided the biological interpretation.

Competing interests

The authors have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable

References

[1] Akhmedov M, Kwee I, and Montemanni R. (2016) A divide and conquer matheuristic algorithm for the Prize-collecting Steiner Tree Problem, *Computers and Operations Research*, **70**, 18-25.

[2] Akhmedov M., LeNail A., Bertoni F., Kwee I., Fraenkel E. and Montemanni R. (2017) A fast Prize-collecting Steiner Forest algorithm for Functional Analyses in Biological Networks, *Lecture Notes in Computer Science*, to appear.

[3] Bechet M.B., Borgs C., Braunstein A., ... and Zecchina R. (2010) Finding undetected protein associations in cell signaling by belief propagation, *PNAS*, **108**, 882-887.

[4] Beisser D., Klau G.W., Dandekar T., Mueller T. and Dittrich M. (2010) BioNet: an R-package for the Functional Analysis of Biological Networks, *Bioinformatics*, **26(8)**, 1129-1130.

[5] Chen E.Y., Tan C.M., Kou Y., .. and Ma'ayan A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, *BMC Bioinformatics*, **128(14)**, PMID:23586463.

[6] Dittrich M.T., Klau G.W., Rosenwald A., ... and Mueller T. (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach, *Bioinformatics*, **24(13)**, i223-i231.

[7] Szklarczyk D., Franceschini A., Kuhn M., Simonovic M., .. and Mering C.v. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, *Nucleic Acids Res*, **39**, D561-D568.

[8] Tan, Y.H.C., Krishnaswamy, S., Nandi, S., Kanteti, R., Vora, S., ... and Robinson, M. (2010) CBL Is Frequently Altered in Lung Cancers: Its Relationship to Mutations in Met and Egfr Tyrosine Kinases, *PloS One*, **5 (1)**, Public Library of Science: e8972.

[9] Thomson S., Petti F., Sujka-Kwok I., Mercado P., ... and Haley, J.D. (2011) A systems view of epithelial-mesenchymal transition signaling states, *Clinical & experimental metastasis*, **28(2)**, PMID:21194007.

[10] Tuncbag N., Gosline S.J.C., Kedaigle A., Soltis A.R., Gitter A. and Fraenkel E. (2016) Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package, *PLoS Comput Biol*, **12(4)**, e1004879.

[11] Tuncbag, N., Braunstein, A., Pagnani, A., Huang, ... and Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of computational biology*, 20(2), 124-136.

[12] Tuncbag, N., McCallum, S., Huang, S. S. C., and Fraenkel, E. (2012). SteinerNet: a web server for integrating 'omic'data to discover hidden components of response pathways. *Nucleic acids research*, gks445.

[13] Care, M. A., Barrans, S., Worrillow, L., Jack, A., Westhead, D. R., and Tooze, R. M. (2013) A microarray platform-independent classification tool for cell of origin class allows comparative analysis of gene expression in diffuse large B-cell lymphoma. *PLoS One*, 8(2), e55895.

[14] Hans, C. P., Weisenburger, D. D., Greiner, T. C., Gascoyne, R. D., Delabie, J., Ott, G., ... and Pan, Z. (2004) Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1), 275-282.

[15] Blenk, S., Engelmann, J., Weniger, M., Schultz, J., ... and Dandekar, T. (2007) Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): analysis of molecular predictors, signatures, cell cycle state and patient survival. *Cancer informatics*, 3.

[16] Margolin, A. A., Nemenman, I., Basso, K., ... and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1), S7.

[17] Johnson, D. S., Minkoff, M., and Phillips, S. (2000) The prize collecting steiner tree problem: theory and practice. *In SODA* (Vol. 1, No. 0.6, p. 4).

[18] Wirapati P. nclust: Fast Hierarchical Clustering. (2010) *R package version 1.7.13*.

[19] Wirapati P. Fast Hierarchical Clustering using Reciprocal Nearest-Neighbor Chain Algorithm. (2009) *SIB Days 2009*

[20] Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[21] Testoni, M., Zucca, E., Young, K. H., and Bertoni, F. (2015) Genetic lesions in diffuse large B-cell lymphomas. *Annals of Oncology*, 26(6), 1069-1080.

[22] Dalla Favera, R. (2017) Molecular Genetics of Aggressive B Cell Lymphoma. *Hematological Oncology*, 35 Suppl 1., in press

[23] Swerdlow, S.H., Campo, E., Pileri, S.A., Harris, N.L., Stein, H., ... and Jaffe, E.S. (2016) The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, 127, 2375-2390.

[24] Ladetto, M., Buske, C., Hutchings, M., Dreyling, M., Gaidano, ... and Zucca, E. (2016) ESMO consensus conference on malignant lymphoma: general perspectives and recommendations for prognostic tools in mature B-cell lymphomas and chronic lymphocytic leukaemia. *Annals of Oncology*, 27, 2149-2160.

[25] Gleeson, M., Jack, A., Cunningham, D., Counsell, N., ... and Barrans, S.L. (2016) The Activated B-Cell Subtype of Diffuse Large B-Cell Lymphoma As Determined By Whole Genome Expression Profiling on Paraffin Embedded Tissue Is Independently Associated with Reduced Overall and Progression Free Survival in the Rituximab Era: Results from the UK NCRI R-CHOP 14 v 21 Phase III Trial. *Blood*, 128, 1746-1746.

[26] Scott, D.W., Mottok, A., Ennishi, D., ... and Gascoyne, R.D. (2015) Prognostic Significance of Diffuse Large B-Cell Lymphoma Cell of Origin Determined by Digital Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue Biopsies. *Journal of Clinical Oncology*, 33, 2848-2856.

[27] Masque-Soler, N., Szczepanowski, M., Kohler, C.W., Spang, R. and Klapper, W. (2013) Molecular classification of mature aggressive B-cell lymphoma using digital multiplexed gene expression on formalin-fixed paraffin-embedded biopsy specimens. *Blood*, 122, 1985-1986.

[28] Veldman-Jones, M.H., Lai, Z., Wappett, M., ... and Thress, K.S. (2015) Reproducible, Quantitative, and Flexible Molecular Subtyping of Clinical DLBCL Samples Using the NanoString nCounter System. *Clinical Cancer Research*, 21, 2367-2378.

[29] Reinholz, M.M., Thompson, D., Botros, I., ... and LaFleur, B. (2016) Next generation sequencing for DLBCL classification. *ASCO Annual Meeting*, Abstract Number: 11559.

[30] Veldman-Jones, M.H., Lai, Z., Wappett, M., Harbron, C.G., Barrett, J.C., Harrington, E.A. and Thress, K.S. (2015) Reproducible, Quantitative, and Flexible Molecular Subtyping of Clinical DLBCL Samples Using the NanoString nCounter System. *Clinical Cancer Research*, 21, 2367-2378.

[31] Ljubic, I., Weiskircher, R., Pferschy, U., Klau, G. W., Mutzel, P., and Fischetti, M. (2006) An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical programming*, 105(2), 427-449.

[32] Cascione, L., Rinaldi, A., Chiappella, A., Kwee, I., ... and Bertoni, F. (2017) Diffuse large B cell lymphoma cell of origin by digital expression profiling in the REAL07 Phase 1-2 study. *British Journal of Haematology*, in press.

[33] Scott, D. W., Wright, G. W., Williams, P. M., ... and Smeland, E. B. (2014). Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*, 123(8), 1214-1217.

[34] Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., and Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *PNAS*, 100(17), 9991-9996.

[35] Izidore S. Lossos, M.D., Debra K. Czerwinski, B.A., ... and Ronald Levy, M.D. (2004) Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes. *N. Engl. J. Med.*, 350:1828-1837.

[36] Bret C., Klein B., Moreaux J. (2012) Gene expression-based risk score in diffuse large B-cell lymphoma *Oncotarget*, 3(12): 1700–1710.

[37] Lenz, G., Wright, G., Dave, S. S., Xiao, W., ... and Vose, J. (2008). Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.*, 359(22), 2313-2323.

[38] Shaknovich, R., Geng, H., Johnson, N. A., ... and Melnick, A. (2010). DNA methylation signatures define molecular subtypes of diffuse large B cell lymphoma. *Blood*, blood-2010.

[39] Klapper, W., Szczepanowski, M., Burkhardt, ... and Hansmann, M. L. (2008). Molecular profiling of pediatric mature B-cell lymphoma treated in population-based prospective clinical trials. *Blood*, 112(4), 1374-1381.

[40] Hummel, M., Bentink, S., Berger, H., Klapper, ... and Hansmann, M. L. (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.*, 354(23), 2419-2430.

[41] Frei, E., Visco, C., Xu-Monette, Z. Y., Dirnhofer, ... and Go, R. S. (2013). Addition of rituximab to chemotherapy overcomes the negative prognostic impact of cyclin E expression in diffuse large B-cell lymphoma. *J Clin Pathol*, 66(11), 956-961.

[42] Williams, P. M., Li, R., Johnson, N. A., ... and Gascoyne, R. D. (2010). A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays. *J Mol Diagn*, 12(5), 680-686.

[43] Salaverria, I., Philipp, C., Oschlies, ... and Berger, H. (2011). Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood*, 118(1), 139-147.