Impacts of dimensionality reduction parameters on Cox Proportional Hazards on cancer gene expressions

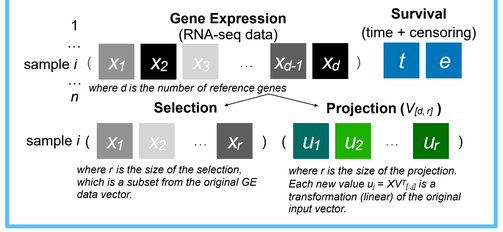
Léonard Sauvé[1], Josée Hébert[1,2,3,4], Guy Sauvageau[1,3,4], Sébastien Lemieux[1,5]

1. Institute for Research in Immunology and Cancer
2. Leukemia Cell Bank of Quebec, Maisonneuve-Rosemont Hospital, Montréal, QC, Canada
3. Division of Hematology-Oncology, Maisonneuve-Rosemont Hospital, Montréal, QC, Canada
4. Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada.
5. Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC, Canada.

INSTITUT DE RECHERCHE EN IMMUNOLOGIE ET EN CANCÉROLOGIE — IRIC — Université de Montréal

Overview

Cancer prognosis at the molecular level from gene expression data (GE) promises to lead to highly accurate models to guide clinical decisions. Cox proportional-hazards (CPH) survival analyses are prone to overfit with high dimensionality input; thus, dimensionality reduction approaches are attractive options. To establish a fair CPH performance baseline as a prerequisite step towards the development of more complex models like the CPH-DNN[1], this work proposes a systematic investigation of the interplay between dimensionality reductions and standard CPH.



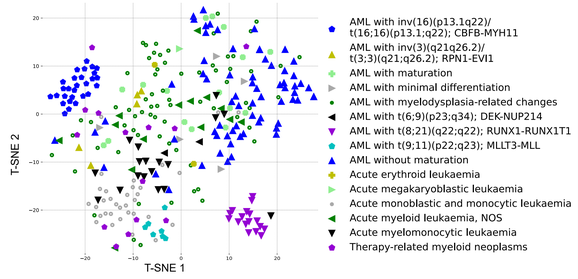
Data

	dataset	input type	nb. samples	input size	censored
input	Leucegene[2]	G	300	19,545	77 (26%)
		G + C	300	19,545 + 8	77 (26%)
	(intermediate)	G	177	19,545	40 (22.6%)
	TCGA[3]	G	140	19,545	55 (39.2%)

G: *gene expression features*, C: *Clinical features*

2-d Vizualization with T-SNE

Leucegene gene expression data (from PCA) by the WHO classification.



Cytogenetic Risk (Leucegene)

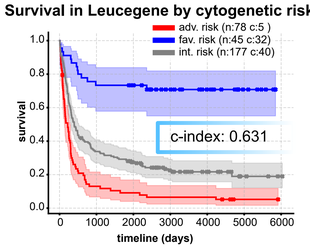
ELN[4] risk stratification by genetics : Risk category & Genetic abnormality

Favorable risk: t(8;21)(q22;q22.1), RUNX1-RUNX1T1. inv(16)(p13.1q22) or t(16;16)(p13.1;q22), CBFB-MYH11. Mutated NPM1 without FLT3-ITD or with FLT3-ITD low. Biallelic mutated CEBPA.

Intermediate risk: Mutated NPM1 and FLT3-ITD high. Wild-type NPM1 without FLT3-ITD or with FLT3-ITD low (without adverse-risk genetic lesions). t(9;11)(p21.3;q23.3), MLLT3-KMT2A. Cytogenetic abnormalities not classified as favorable or adverse.

Adverse risk: t(6;9)(p23;q34.1), DEK-NUP214. t(v;11q23.3), KMT2A rearranged. t(9;22)(q34.1;q11.2), BCR-ABL1. inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2), GATA2, MECOM(EVI1). -5 or del(5q). -7;-17/abn(17p), ...

Survival in Leucegene by cytogenetic risk

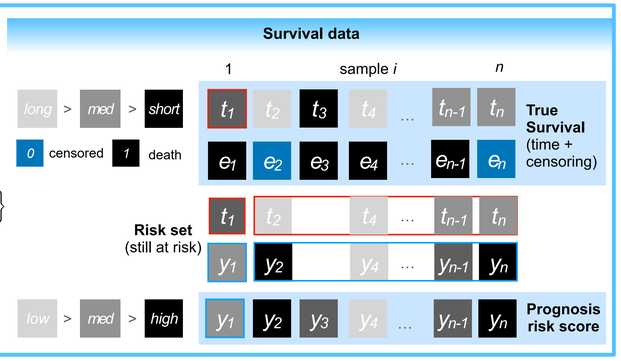


Cox Proportional Hazards[5]

survival function $\lambda(t;x) = \lambda_0(t)\exp(x^T\beta)$

Loss function $\ell(\beta) = \sum_{r\in D}\left\{x_r^T\beta - \log\left(\sum_{i\in R_r}\exp(x_i^T\beta)\right)\right\}$

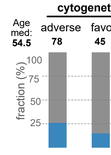
concordance index (c index) $C = P(\widehat{T}_1 > \widehat{T}_2 | T_1 > T_2)$



Validation of sample reclassification and correlations between GE to clinical features (mutations, cyto-risk, sex and age)

Goal: show usefulness of GE for prognosis in AML.

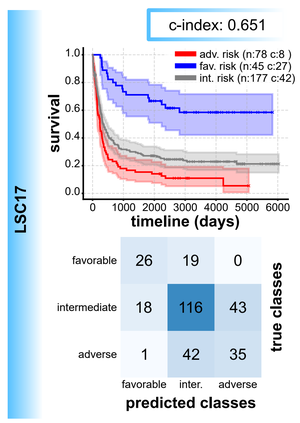
Methods: We used the Leucegene dataset to compare the cytogenetic risk to the PCA and LSC17-CPH models' 3-way stratification. Survival curves with *lifelines*. The sample size for each category is set to the original distribution of the data (fav: 45, int: 177, adv: 78). The confusion matrices were obtained using *scikit-learn*. *Pearson-moment correlations* were computed between individual genes / PC and clinical factors using *scikit-learn*.



Results: LSC17 presents fair similarities to cytogenetic risk classification, although LSC17 reclassifies many samples to adjacent risk categories. Some of the LSC17 genes have strong correlations with the NPM1 mutation and the favorable cytogenetic risk.

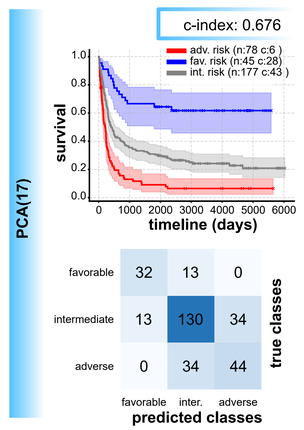
PCA(17) as with LSC17, PCA risk classification is faithful to the cytogenetic risk classification, although it reclassifies many samples to adjacent risk categories. Some components correlate to risk and mutations (but less than LSC17).

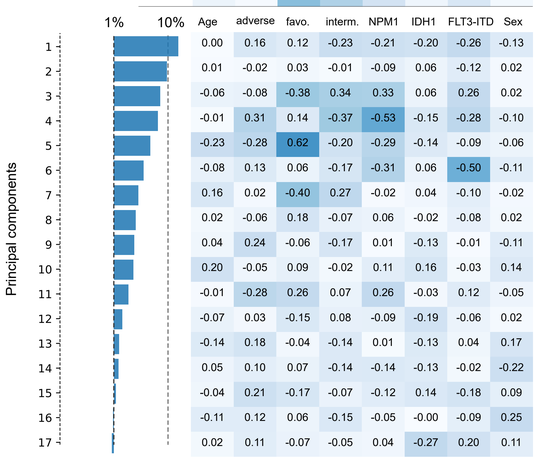
LSC17



LSC17 genes	Age	adverse	favo.	interm.	NPM1	IDH1	FLT3-ITD	Sex
CD34	-0.07	0.13	0.41	-0.42	-0.59	-0.17	-0.28	-0.05
LAPTM4B	0.11	0.13	-0.32	0.11	0.20	-0.05	0.32	0.06
ARHGAP22	0.09	-0.02	-0.32	0.25	0.51	0.18	0.47	0.21
LOC284422	-0.04	-0.11	0.17	-0.02	-0.06	0.11	0.06	0.13
KIAA0125	0.04	-0.09	0.35	-0.17	-0.15	-0.06	-0.10	0.05
MMRN1	-0.07	0.22	-0.05	-0.16	-0.21	-0.06	0.04	0.06
DNMT3B	0.07	0.11	-0.56	0.31	0.26	0.15	0.31	0.13
SOCS2	-0.01	0.15	-0.10	-0.06	-0.01	-0.11	0.39	0.04
CDK6	-0.11	-0.14	-0.06	0.17	0.04	0.18	0.10	0.12
NGFRAP1	0.00	0.13	-0.07	-0.06	0.07	0.05	0.03	0.01
CPXM1	-0.01	-0.22	0.13	0.11	0.14	0.07	0.30	0.10
ZBTB46	-0.03	-0.02	-0.11	0.10	0.12	0.10	0.17	0.08
DPYSL3	0.11	-0.05	-0.13	0.14	0.26	0.22	0.17	0.15
NYNRIN	0.03	0.17	-0.40	0.14	0.06	-0.01	0.24	0.08
GPR56	0.05	-0.03	-0.28	0.23	0.30	0.09	0.33	0.15
AKR1C3	0.04	-0.09	0.16	-0.03	-0.08	0.05	0.02	0.08
EMP1	-0.06	0.01	0.43	-0.32	-0.19	0.05	-0.16	-0.02

PCA(17)



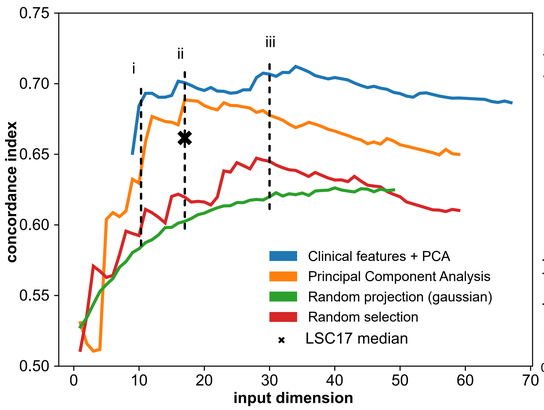


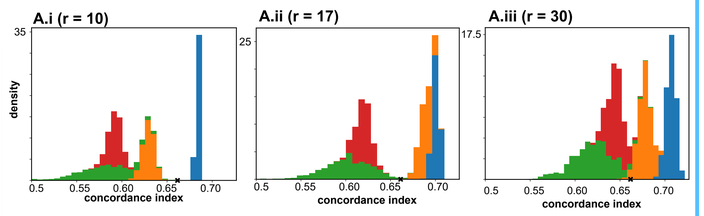
PCA(17)	Age	adverse	favo.	interm.	NPM1	IDH1	FLT3-ITD	Sex
1	0.00	0.16	0.12	-0.23	-0.21	-0.20	-0.26	-0.13
2	0.01	-0.02	0.03	-0.01	-0.09	0.06	-0.12	0.02
3	-0.06	-0.08	-0.38	0.34	0.33	0.06	0.26	0.02
4	-0.01	0.31	0.14	-0.37	-0.53	-0.15	-0.28	-0.10
5	-0.23	-0.28	0.62	-0.20	-0.29	-0.14	-0.09	-0.06
6	-0.08	0.13	0.06	-0.17	-0.31	0.06	-0.50	-0.11
7	0.16	0.02	-0.40	0.27	-0.02	0.04	-0.10	-0.02
8	0.02	-0.06	0.18	-0.07	0.06	-0.02	-0.08	0.02
9	0.04	0.24	-0.06	-0.17	0.01	-0.13	-0.01	-0.11
10	0.20	-0.05	0.09	-0.02	0.11	0.16	-0.03	0.14
11	-0.01	-0.28	0.26	0.07	0.26	-0.03	0.12	-0.05
12	-0.07	0.03	-0.15	0.08	-0.09	-0.19	-0.06	0.02
13	-0.14	0.18	-0.04	-0.14	0.01	-0.13	0.04	0.17
14	0.05	0.10	0.07	-0.14	-0.14	-0.13	-0.02	-0.22
15	-0.04	0.21	-0.17	-0.07	-0.12	0.14	-0.18	0.09
16	-0.11	0.12	0.06	-0.15	-0.05	-0.00	-0.09	0.25
17	0.02	0.11	-0.07	-0.05	0.04	-0.27	0.20	0.11

Determination of adequate dimensionality reduction policy for standard CPH training on cancer data

Methods: Principal component analysis (PCA), random projection, random selection, and **published gene signature LSC17[6]** were evaluated through 5-fold cross-validation on the Leucegene whole cohort, intermediate and on the TCGA dataset. Only the 50% most expressed genes were selected. We reported the bootstrapped risk scores (n=1000) on the validation set to get the distribution of c indices. Models were trained using inputs of increasing dimensions. All models are implemented in python, via the *scikit-learn* and *lifelines* packages. Survival curves were drawn with lifelines Kaplan-Meier fitting method and number of censored samples were reported for each predicted risk group. L2 regularization parameter was set to 0.001 with no L1.

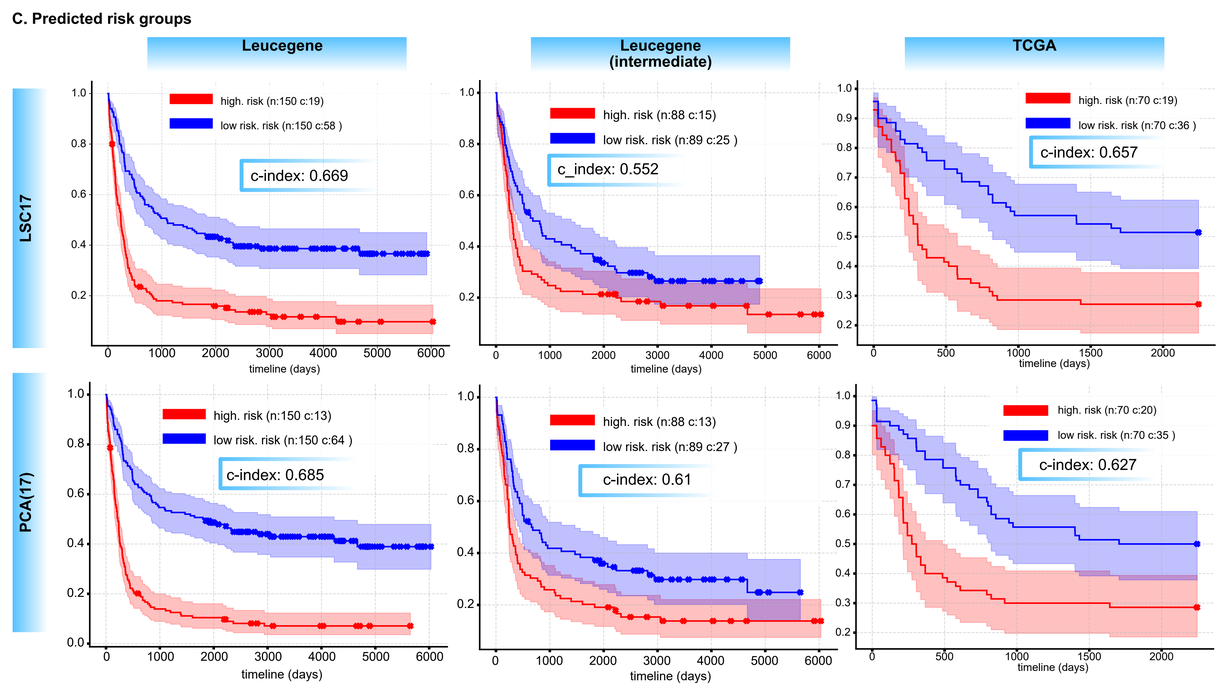
A. Performance of **Cox-PH** model from increasing dimensionality inputs





Results: (A) Multivariate model (PCA+CF) is the model with highest accuracy overall. At equal dimensions (17), PCA gets better concordance than LSC17. **(A.iii)** LSC17 performs among the best 30 random genes signatures. **(B)** CPH models suffer from overfitting with increasing dimensionality inputs. **(C)** PCA gets better concordance (0.685) than the cytogenetic risk baseline (see Data, c index = 0.631) and performs as well (or better) than the published LSC17 gene signature (c = 0.669) in Leucegene (whole cohort or intermediate subset). It is not clear that PCA gets better concordance than LSC17 in the TCGA dataset.

C. Predicted risk groups



Discussion / perspectives

PCA gets equal or better concordance and **requires less *a priori* knowledge** than gene signatures such as LSC17. Simpler by design, PCA might be better suited for cancer prognosis from GE data. The discordance between **PCA / LSC17** with **cytogenetic risk** on sample risk stratification (but getting better concordance) means using GE for clinical molecular prognosis could lead to more accurate risk assessment of the intermediate samples. PCA / LSC17 individual components correlate to already in use clinical factors. The question remains whether gene expressions merely captures standard factors or if the profiles contain yet unknown useful survival information as well.

References

1. **Faraggi, D., & Simon**, R. (1995). **A neural network model for survival data.** *Statistics in Medicine*, *14*(1), 73–82.
2. **www.leucegene.ca**
3. **https://www.cancer.gov/tcga**
4. **Döhner, H.**, Estey, E., … Bloomfield, C. D. (2017). Diagnosis and management of AML in adults: 2017 **ELN recommendations from an international expert panel.** *Blood*, *129*(4), 424–447.
5. **Hao, L.**, Kim, J., Kwon, S., & Ha, I. D. (2021). **Deep Learning-Based Survival Analysis for High-Dimensional Survival Data.** *Mathematics*, *9*(11), 1244.
6. **Ng, S.** W. K., Mitchell, A., Kennedy, J. A., … Wang, J. C. Y. (2016). **A 17-gene stemness score for rapid determination of risk in acute leukaemia.** *Nature*, *540*(7633), 433–437.