Splicing-aware Patient Stratification

Yazgi Sert
Technical University of Munich

Introduction

- Complex diseases like cancer, cardiovascular diseases, and asthma are highly heterogeneous.
- In heterogeneous diseases, two patients with the same diagnosis might have different disease mechanisms.
- Patients can be stratified into subtypes based on similar clinical measurements and omics profiling patterns.
- Patient subtypes help in addressing disease heterogeneity.
- Clustering and Biclustering address patient subtyping in an unsupervised manner.
- Biclustering techniques perform simultaneous clustering of patients and measurements.
- Biclustering is usually done at gene level. However, gene expression is a sum of transcript expressions due to alternative splicing.
- Biclustering simultaneously clusters patients and parameters (e.g., gene expression) and aims to detect disease subtypes.
- Biclustering is suited for identifying disease-associated genes from gene expression data while stratifying patients at the same time.

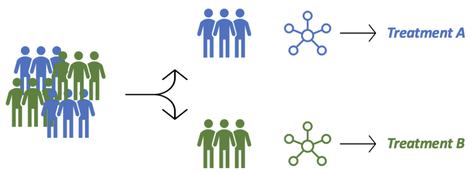

Fig. 1: Patient stratification

Motivation

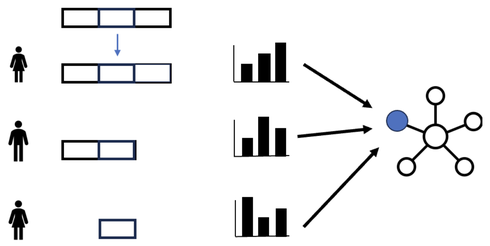

Fig. 2: Biclustering at transcript level

- Biclustering is usually done at gene level. Gene expression is sum of transcript expressions. Biclustering at transcript level could uncover clinically relevant subtypes in complex diseases.
- Our aim: To investigate the feasibility of splicing-aware measurements for patient stratification for complex diseases
- Mechanisms of some complex diseases involve dysregulation of splicing.

Used Data

- Expression file: breast cancer RNA-Seq data set from the Cancer Genome Atlas (TCGA)
- Network: BioGRID, The Biological General Repository for Interaction Datasets

Methods

The network-constrained methods aim to extract functionally relevant biclusters using molecular interaction network as constraints: the bicluster is significant if it is enriched with proteins that are neighbors in the molecular interaction network.

Methods

BiCoN:

- Accepts gene expression data as input and stratifies patients into two subgroups while identifying, for each group, a subnetwork of genes that can be interpreted as a shared molecular mechanism.
- Extracts a fixed number of non-overlapping biclusters, which are connected in a molecular interaction network.
- Resulting subnetwork can be interpreted as a biological function jointly carried out by these genes which is active in one patient group and inactive in the other one.
- Leverages molecular interaction networks in the analysis of gene expression data to detect known subtypes as well as novel, clinically relevant patient subgroups [1].

DESMOND:

- Accepts matrix of normalised gene expression data and network of gene interactions as input.
- Identifies gene modules - connected sets of genes up- or down-regulated in subsets of samples.
- Allows identifying the biologically meaningful gene and sample subsets and improves the reproducibility of the results.
- Applies flexible thresholds for identification of samples in which genes are differentially expressed.
- Problem: discovery of connected groups of genes differentially expressed in an unknown subgroup of samples, given a network of gene interactions and a matrix of gene expression profiles [2].

Results

These are the example cluster map and network results from one run on gene level and one on transcript level.

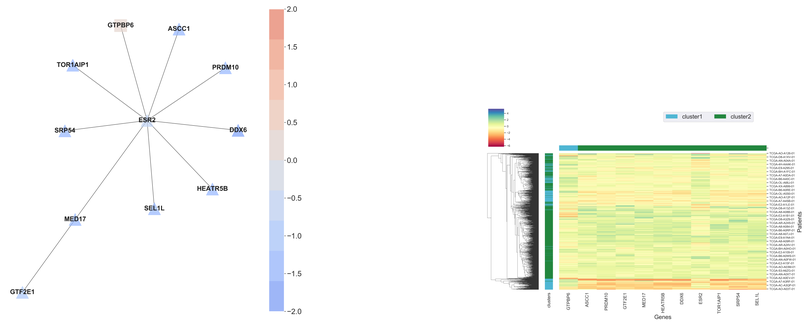

Fig. 3: Cluster map and network results from BiCoN (on gene level)

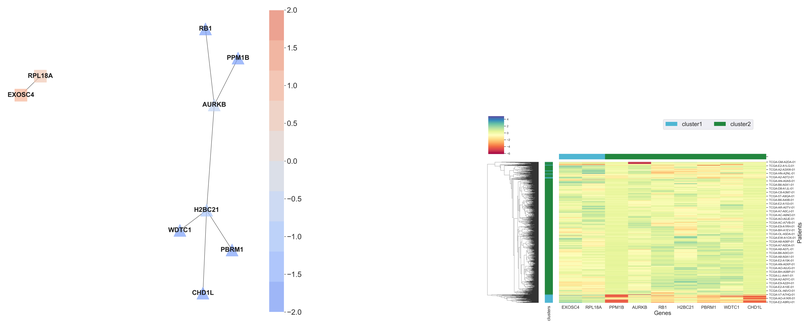

Fig. 4: Cluster map and network results from BiCoN (on transcript level)

For transcript level data, we are considering the most dominant isoform by median expression.

Gene	Frequency in runs
PDRM1O	4
MED17	4
HEATR5B	4
DDX6	4
ESR2	4

From gene level data

Gene	Frequency in runs
CAPZA1	2
NOLC1	2
RRS1	2
PRPF8	2
VHL	2

From transcript level data

Fig. 5: Frequencies of genes in 7 runs

Results

For Box-Plotting, we considered the maximum Jaccard index of each subtype present in the clusters.
A high Jaccard Index indicates a high degree of similarity between the cluster and the subtype.

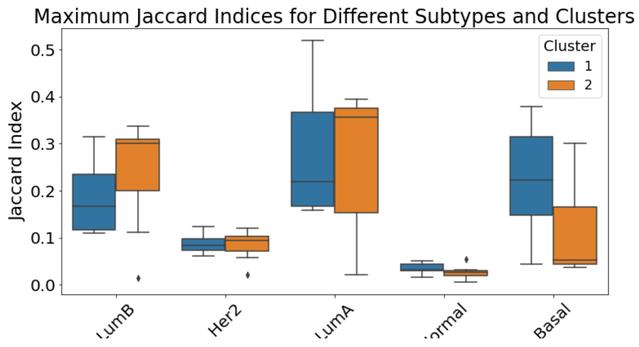

Fig. 6: Box-Plot for Biclustering on gene level

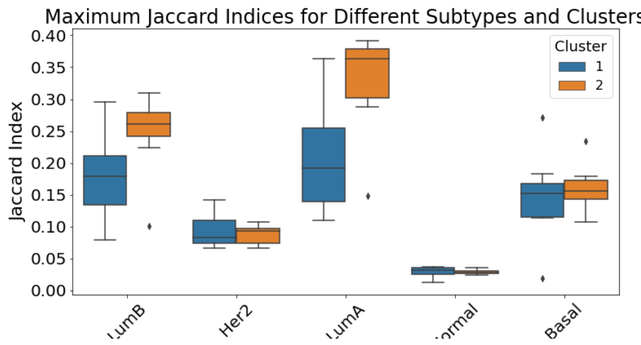

Fig. 7: Box-Plot for Biclustering on transcript level

Conclusion

- Transcript level data works worse than gene level data. This probably relies on not using the appropriate measurement, while extracting information from the expression file on transcript level.
- In some cases, one of the resulting clusters include only one or even zero genes.
- The transcript results are not robust, as we can see from the difference between the network result and the occurence frequencies of the genes throughtout the runs.

Next Steps

- Compare DESMOND with BiCoN.
- Extend network-constrained biclustering approaches for splicing-aware stratification.
- Implement several measurements that would reflect the level of splicing of a gene.
- Investigate the applicability of these measurements for splicing-aware patient stratification based on network-constrained biclustering approaches, implemented in DESMOND and BiCoN.

References

[1] Lazareva, O., Van Do, H., Canzar, S., Yuan, K., Tieri, P.,Baumbach, J., Kacprowski, T., List, M.: BiCoN: Network-constrained biclustering of patients and omics data. [Submitted]

[2] Zolotareva, O., Khakabimamaghani, S., Isaeva, O. I., Chervontseva, Z., Savchik, A.,& Ester, M. (2020). Identification of differentially expressed gene modules in heterogeneous diseases. Bioinformatics. Oxford University Press. https://doi.org/10.1093/bioinformatics/btaa1038