NEVADA NASA SPACE GRANT CONSORTIUM

DSCC: DISEASE SUBTYPING USING COMMUNITY DETECTION FROM CONSENSUS NETWORK

UNIVERSITY OF NEVADA, RENO

Hung Nguyen, Monikrishna Roy, Ha Nguyen, and Tin Nguyen*

Department of Computer Science and Engineering, University of Nevada, Reno
Contact: tinn@unr.edu. Website: https://bioinformatics.cse.unr.edu

04/30/2021

MOTIVATION

Cancer subtyping is crucial to improve treatment and prognosis. Multi-omics data integration is important because it allows us to differentiate among subtypes from a holistic perspective that takes into consideration phenomena at various levels (proteomics, mutations, etc.). However, the following challenges need to be overcome:

- Missing data (e.g., a patient has mRNA but not methylation)
- The integration of continuous and categorical variables
- High-dimensionality and large sample sizes

OBJECTIVES

Develop a robust disease subtyping method that is able to (1) handle missing data, (2) integrate continuous and categorical data, and (3) cope with big data scale (large sample sizes and high-dimensionality).

RESULTS

Data: 33 cancer datasets contain 10 data types with a total of 11,085 samples and clinical data for each patient from Genomic Data Common Portal (https://portal.gdc.cancer.gov/).

Metric: Cox p-value that measures the significance in survival difference of the discovered subtypes.

Methods: DSCC, NEMO [3], SNF [4], CIMLR [5], CC [6], and LRACluster [7] .

Results: DSCC was able to (1) discover novel subtypes with significantly different survival profiles in most datasets (18 out of 33 datasets), and (2) has the most significant p-values (highest minus log p-values).

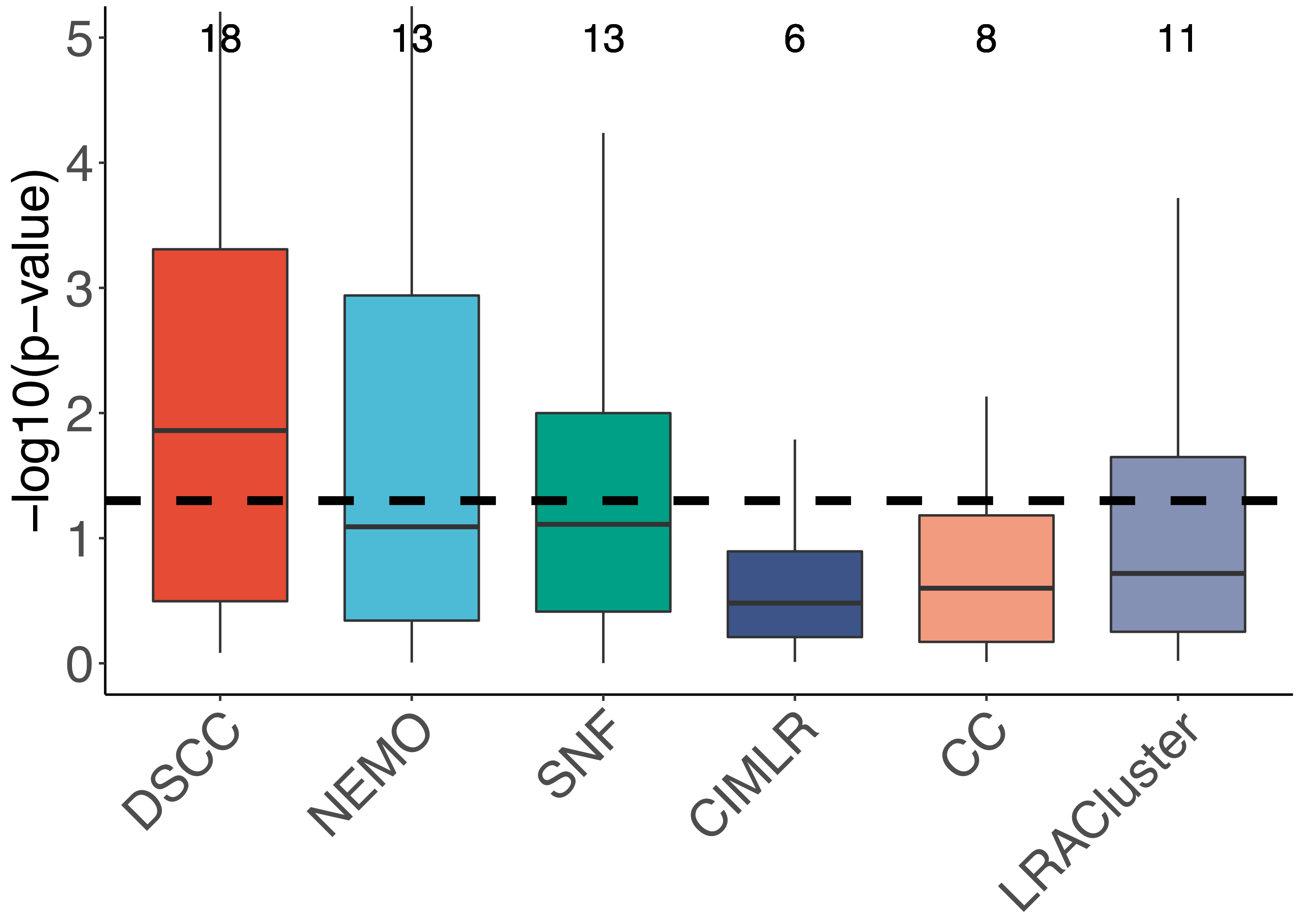

***Figure 1:** Analysis results using 33 cancer datasets.*

DSCC METHOD

(i) Inputs: Multi-omics data of cancer patients, and cancer pathways available on Kyoto Encyclopedia of Genes and Genomes [1].

(ii) Dimension Reduction: (1) Project data into pathways; (2) Perform factor analysis for continuous data and multiple correspondence analysis (MCA) for categorical data; and (3) Perform principal component analysis.

(iii) Patient network construction: (1) Build connectivity matrix for each data type using consensus clustering; and (2) Combine similarity matrices using Kolmogorov–Smirnov test (KS test) and compatibility metrics.

(iv) Subtyping: Partition the patient network using a community detection algorithm (Louvain [2]).

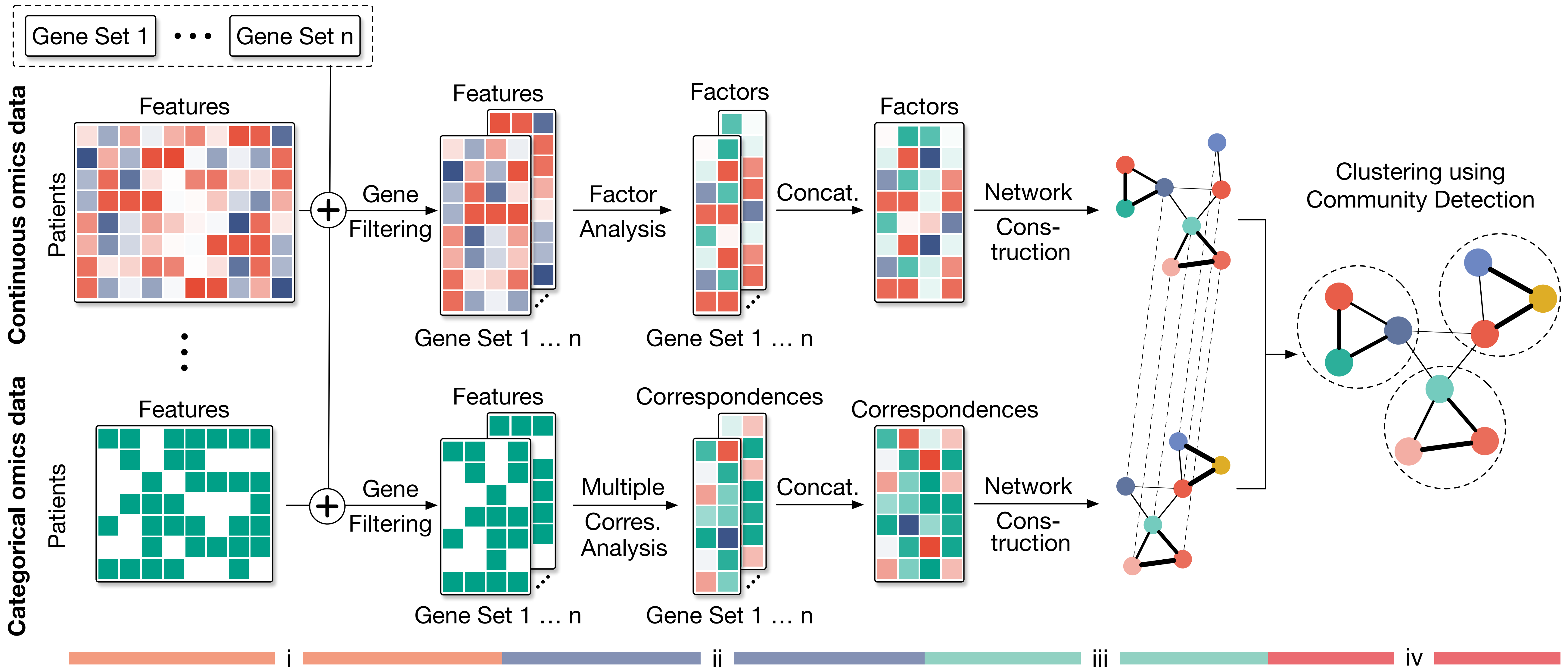

***Figure 2:** The DSCC computational framework for multi-omics integration and cancer subtyping.*

CONCLUSION

Analysis results on 33 cancer datasets demonstrate that DSCC is able to identify novel subtypes with significantly different survival profiles. The approach can integrate both numerical and categorical data. Another important property of DSCC is that it is able to analyze data with missing values, i.e., not all patients have all data types measured.

ACKNOWLEDGEMENT

This work was partially supported by NASA under grant numbers 80NSSC19M0170 and NNX15AI02H (subaward no. 21-02), and by NIH NIGMS under grant number GM103440

References

1. Kanehisa et al., "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Research 44.D1 (2016), pp.D457-D462.
2. Blondel et al., "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment , 2008.10 (2008): P10008.
3. Rappoport and Shamir., "NEMO: Cancer subtyping by integration of partial multi-omic data.", Bioinformatics, 35.18 (2019): 3348-3356.
4. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale.", Nature Methods, 11.3 (2014): 333.
5. Ramazzotti et al., "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival.", Nature Communications, 9.1 (2018): 1-14.
6. Monti et al., "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data." Machine Learning, 52.1 (2003): 91-118.
7. Wu et al., "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification.", BMC Genomics 16.1 (2015): 1022.
8. Genomic Data Commons Data Portal, https://portal.gdc.cancer.gov.