॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥ Department of Bioscience & Bioengineering

ICGA FOUNDATION | Indian Cancer Genome Atlas

3rd ICGA Conference : 13th-14th January 2022
Biobanking to Omics: Collecting the Global Experience

Integrating Multi-omics Data for Cancer Subtyping: A Multi-view Clustering Algorithm

Madhumita, Archit Dwivedi, and Sushmita Paul
Department of Bioscience and Bioengineering
Indian Institute of Technology Jodhpur, Rajasthan, India

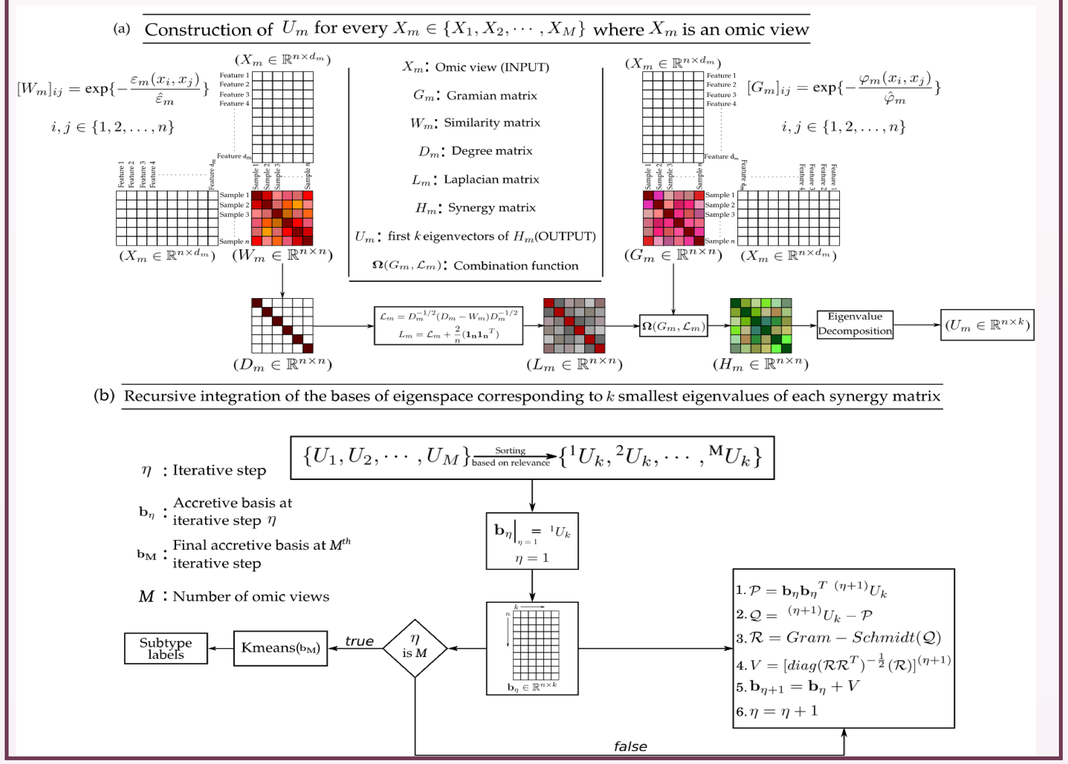
ABSTRACT

Cancer subtypes identification is one of the critical steps towards advancing personalized anti-cancerous therapies. It can provide deeper insights into the molecular signatures targeted to understand disease pathogenesis. Accumulation of a considerable amount of multi-platform omics data, like, Transcriptome, Proteome, Epigenome, and others measured on the same set of samples, provides an opportunity to look into this deadly disease from several views simultaneously. A novel multi-view clustering algorithm named RISynG is presented in this study that integrates multi-omics data and performs clustering across the samples for cancer subtypes identification. RISynG is tested on five multi-omics cancer data sets taken from The Cancer Genome Atlas. The experimental results demonstrate that the proposed method outperforms other methods in this domain. The cancer subtypes identified by RISynG overlaps with the well-established and studied subtypes of respective cancer types to a greater extent.

CONTACT

Madhumita
Department of Bioscience and Bioengineering,
IIT Jodhpur, Rajasthan
Email: madhumita.1@iitj.ac.in
Phone: +91-6377583996
Website:https://home.iitj.ac.in/~madhumita.1/Madhumita.html

Proposed Muli-omics Clustering Algorithm: RISynG (Recursive Integration of Synergized Graph-representations)



CONCLUSION

The main contributions of this study are:

1) **Development of an integrative clustering method for multi-view Omics data.**
2) **Demonstration of the effectiveness of proposed method over other methods.**
3) **Establishing biological relevance for the obtained results.**

REFERENCES

1. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, ser. NIPS'01. Cambridge, MA, USA: MIT Press, 2001, p. 849–856
2. W. Bo, M. M. Aziz, D. Feyyaz, F. Marc, T. Zhuowen, B. Michael, H. K. Benjamin, and G. Anna, "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale," Nature Methods, vol. 11, no. 3, pp. 333–337, 2014.
3. A. Khan and P. Maji, "Selective Update of Relevant Eigenspaces for Integrative Clustering of Multimodal Data," IEEE Transactions on Cybernetics, pp. 1–13, 2020.
4. The Cancer Genome Atlas Network, "Integrated genomic and molecular characterization of cervical cancers," Nature, vol. 543, no. 7645, pp. 378–384, 2017.
5. S. Paul and Madhumita, "RFCM3: Computational Method for Identification of miRNA-mRNA Regulatory Modules in Cervical Cancer," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1729–1740, 2020.
6. S. M. Agarwal, D. Raghav, H. Singh, and G. Raghava, "CCDB: A Curated Database of Genes Involved in Cervix Cancer," Nucleic Acids Research, vol. 39, no. 1, pp. D975–D979, 2011.

ACKNOWLEDGEMENTS

This work is partially supported by the seed grant program of the Indian Institute of Technology Jodhpur, India (grant no. I/SEED/SPU/20160010). The authors acknowledge Dr. Sukhendu Ghosh, Department of Mathematics, Indian Institute of Technology Jodhpur for his fruitful discussions.

DATA SETS

		Number of features				
Data-sets	Number of samples	mRNA	miRNA	metDNA	RPPA	Number of clusters
CESC	124	2000	311	2000	219	3
BRCA	398	2000	278	2000	212	4
OV	474	2000	591	-	-	2
LGG	267	2000	333	2000	209	3
STAD	223	2000	524	-	-	4

RESULTS

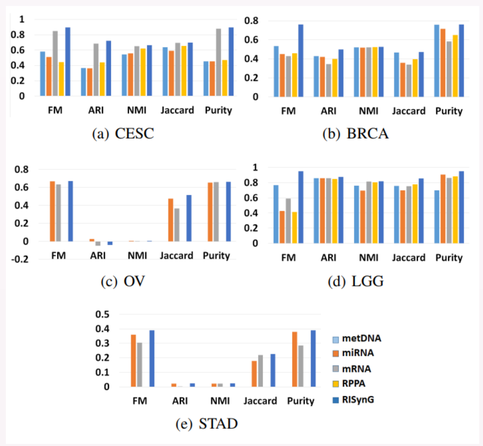


Fig 1: Comparative performance analysis of proposed approach and spectral clustering performed on individual omic-views

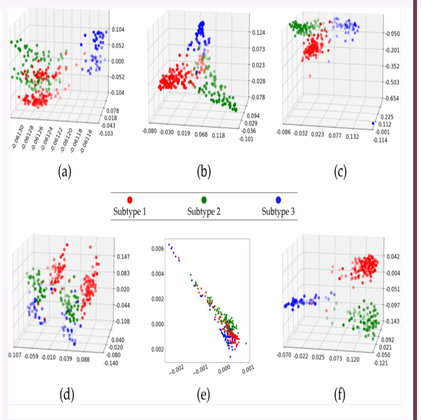


Fig 2: Comparative analysis of different integrative sub-spaces on LGG data set. (a) Best-omic view (metDNA) (b) SNF (c) SURE (d) CoALa (e) iCluster (f) RISynG

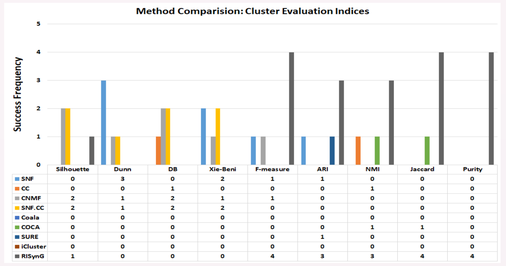
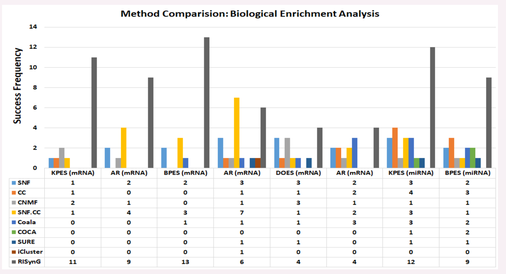


Fig 3 &4: Comparative Analyses: Success frequency indicates number of times a method scored the maximum value for each of the evaluation indices.



Methods		yes	no	Total	*p*-value
SNF	yes	22	202	224	0.715
	no	163	1613	1776	
CC	yes	20	198	218	1.000
	no	165	1617	1782	
CNMF	yes	26	197	223	0.219
	no	159	1618	1777	
SNF.CC	yes	22	202	224	0.715
	no	163	1613	1776	
CoALa	yes	19	200	219	0.902
	no	166	1615	1781	
COCA	yes	23	194	217	0.457
	no	162	1621	1783	
SURE	yes	18	207	225	0.543
	no	167	1608	1775	
iCluster	yes	17	205	222	0.460
	no	168	1610	1778	
RISynG	yes	30	192	222	0.026
	no	155	1623	1778	
Total		185	1815	2000	

Fig 5: Overlap Analysis for Cervical cancer: Overlap between experimentally validated genes and differentially expressed genes obtained between identified subtypes using Fisher's exact test .