Integrative analyses in omics data: Machine learning perspective

Integrative Analysen von Omics-Daten: Perspektive des maschinellen Lernens

Abstract

Developments in the high throughput technologies have enabled the production of an immense amount of knowledge at the multi-omics level. Considering complex diseases which are affected by multi-factors, single omics datasets might not be sufficient to unveil the molecular mechanisms of heterogeneous diseases. Providing a comprehensive and systematic overview to explain disease hallmarks in significant depth is critical. Utilizing multi-omics datasets has led to the development of a variety of tools and platforms. Machine learning models are utilized in a wide variety of tools to tackle the complexity of disorders and to identify new biomolecular signatures and potential markers. Underlying aspects of these approaches are based on training the models for making predictions and classification of the given data. In this review, we describe current machine learning-based approaches and available implementations. Challenges in the enlightenment of disease mechanisms of onset and progression and future development of the field of medicine will be discussed. The prominence of biological interpretation of model output with corresponding biological knowledge will be also covered in this review.

Zusammenfassung

Die Entwicklungen im Bereich der Hochdurchsatztechnologien haben den Erwerb einer immensen Menge an Wissen auf der Multi-Omics-Ebene ermöglicht. In Anbetracht komplexer Krankheiten, die von mehreren Faktoren beeinflusst werden, reichen einzelne Omics-Datensätze möglicherweise nicht aus, um die molekularen Mechanismen heterogener Krankheiten aufzudecken. Ein umfassender und systematischer Überblick ist notwendig, um Krankheitsmerkmale ausreichend zu erklären. Die Verwendung von Multi-Omics-Datensätzen hat zur Entwicklung einer Vielzahl von Werkzeugen und Plattformen geführt. Modelle des maschinellen Lernens werden in einer Vielzahl von Instrumenten eingesetzt, um die Komplexität von Krankheiten zu erfassen und neue biomolekulare Signaturen und potenzielle Marker zu identifizieren. Die grundlegenden Aspekte dieser Ansätze beruhen auf dem Training der Modelle, um Vorhersagen und Klassifizierungen der gegebenen Daten vorzunehmen. In dieser Übersichtsarbeit beschreiben wir die aktuellen, auf maschinellem Lernen basierenden Ansätze und die verfügbaren Implementierungen. Die Herausforderungen bei der Aufklärung der Mechanismen von Krankheitsentstehung und Krankheitsverlauf und zukünftige Entwicklungen im Bereich der Medizin werden erörtert. Auch die Bedeutung der biologischen Interpretation von Modellergebnissen mit entsprechendem biologischen Wissen wird in dieser Übersichtsarbeit angesprochen.

Miray Unlu Yazici[1]
Burcu Bakir-Gungor[2]
Malik Yousef[3,4]

1 Department of Bioengineering, Faculty of Engineering, Abdullah Gül University, Kayseri, Turkey

2 Department of Computer Engineering, Faculty of Engineering, Abdullah Gül University, Kayseri, Turkey

3 Department of Information Systems, Zefat Academic College, Zefat, Israel

4 Galilee Digital Health Research Center, Zefat Academic College, Zefat, Israel

Overview of omics data types

Collective characterization and quantification of biomolecules with advanced technologies have yielded the study of fields such as genome, transcriptome, epigenome, metabolome, etc. Initiation of omics studies with genomics lead to early diagnosis and target treatments via understanding the mechanisms of diseases. Genomics driven genetic variations on phenotype are analyzed with different methods and databases, such as the genome-wide association study (GWAS) [1] and Gene Expression Omnibus (GEO) [2]. Transcriptomics data publicly available in GEO and Sequence Read Archive (SRA) [3] enable the identification of novel transcripts and expression value of transcripts in RNA level studies. The PRoteomics IDEntifications (PRIDE) [4] and ProteomicsDB [5] profile mass spectrometry-based proteome changes. Furthermore, whole exome sequencing (WES) studies focus on protein coding regions of genes to identify genetic variants affecting the mechanism of diseases. The Genome Aggregation Database (gnomAD) [6] provides whole genome and exome sequencing data from large-scale sequencing projects. Interactome provides molecular interaction wiring in cells. The interactome databases such as IntAct [7], BioGrid [8], and STRING [9] are utilized to understand the dynamic interplay of molecules in developing novel therapeutic strategies. For instance, cross-link with neighboring proteins can lead to a basis for their role in signaling pathways and identification of molecular targets of specific drugs.

Genetic changes rewire the cellular networks in complex diseases. Multi-omics data obtained from the same set of samples can enlighten the mechanisms underlying the disease heterogeneity via detecting more coherent signatures and relevant interactions through flow of genetic information. The publicly available repositories The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/), International Cancer Genomics Consortium (ICGC, https://icgc.org/), and Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle) provide several types of multi-omics data in cancer. While the Therapeutically Applicable Research To Generate Effective Treatments (TARGET, https://ocg.cancer.gov/programs/target) database includes pediatric cancer-related omics data at the biological level, the datasets of human, model and non-model organisms can be accessed from the repository Omics Discovery Index (OmicsDI, https://www.omicsdi.org/).

Machine learning perspective in omics data analysis

The analyses of pattern recognition and making predictions based on high dimensional omics data has enabled the machine learning models to capture the patterns accurately compared to traditional mathematical models. Supervised learning models are trained with labeled data and the evaluated model is used for prediction. Unsupervised learning models identify hidden patterns in unlabeled data.

Unsupervised learning mostly covers dimension reduction techniques and association analyses. Clustering-based unsupervised integration method is used to identify disease and molecular subtypes and grouping of features. The Similarity Network Fusion approach (unsupervised) creates sample-sample similarity matrix for each omics data type and merges the matrices [10]. Network based unsupervised integration approaches are based on statistical models and functional interactions of features. In the constructed network, edges represent the predicted relationships of different signatures (nodes) such as genes, CpGs and proteins [11].

Several approaches have been collected under the umbrella term of supervised learning. Support vector machine algorithms (SVM) classify the features by finding hyper-planes. Meta-analytic SVM allows multiple omics data analysis and potential biomarker detection for integrating multiple omics data [12]. The k-Nearest Neighbor (kNN) algorithm based on distance-based method uses a feature similarity approach to calculate the distance from all features around the unknown data to predict the class of it. The kNN Graph (kNN-G), which is widely used in single cell analysis, detects communities or clusters of related cells based on, for example, gene expression data and RNA-Seq profiles [13]. Random forest algorithm based on building random decision trees uses bootstrap aggregation method for class prediction in classification tasks. Random forest with the components recursive feature elimination and permutation-based feature selection providing significance label for the selected feature is used in omics data analysis for the diagnosis of the diseases [14].

Most of the feature selection methods in ML perform omics data analyses with statistics and computer science, called as fully data driven approaches, disregarding biological domain knowledge. The domain knowledge such as disease-gene, drug-disease associations, and protein-protein interactions is entitled as pre-existing biological knowledge. In the following part, we will discuss the studies including pre-existing, fully data driven or a combination of them.

Integrative approach by utilizing pre-existing biological knowledge

Biological systems are massively complex and heterogeneous in nature. To understand the processes holistically in complex organisms, the interpretation of biological data generated in massive volume via high throughput technologies is imperative. Integrating omics data types and utilizing the flow of information among them have facilitated researchers to decipher the field of medicine and biology. Constructing a framework on multi-dimensional biological data integration such as clustering and machine learning approaches can provide a comprehensive understanding of the biological mechanisms under study.

A cost-related limited number of samples for omics data generation is a challenge. The phenomenon, curse of dimensionality, reported by Bellman et al. defines this kind of obstacle with data in high-dimensional spaces. Dimension of the gene or biological features with functional metrics are crucial for prediction, optimization problems and performance results in machine learning (ML).

The assessment of gene expression to unveil the relationship between genotype and phenotype has led scientists to advance in novel methodologies such as DNA microarray and RNA-seq. Previous conventional studies include standard ML and clustering procedures [15], [16], [17] for biomarker discovery [18]. The immense amount of biological knowledge has deflected the course of action of studies from pure data-oriented to integration-based approaches. The advanced tools, platforms, and software developed by bioinformaticians have incorporated biological knowledge into the knowledge base and improved the performance analysis of biological processes. Some of the organized biological knowledge in databases are miRTarBase [19] identifying miRNA-target interactions, Gene Ontology (GO) [20] describing the attributes of genes, KEGG pathways providing molecular interaction networks [21], and DisGeNET [22] targeting disease-gene associations.

Conventional feature selection algorithms typically performed in gene expression analysis rely on statistical and machine learning models. Improving the models by integrating the biological knowledge can contribute to better performance. Current approaches used in gene-expression analyses are reviewed in [23].The authors surveyed the clustering methods with several distance measures such as Euclidean and Manhattan distance, Kendall, and Pearson correlations. Biological background information from external sources [24] and statistics provided to integrative gene selection approaches are used in the identification of informative genes. In this context, the conducted studies aim to improve the classification performance, and biological relevance of significant genes. Gene Ontology (GO), one of the extensively used external sources, exploits the domain knowledge and yields computable gene knowledge by defining classes of gene functions. The Gene Ontology Consortium summarized the studies incorporating GO into statistical analysis to reveal GO terms associated with given genes [25]. Liang et al. presented the enrichment analysis of differentially expressed genes by capturing significant KEGG pathways with a modified Fisher's exact test [12]. Another study conducted by Wang et al. introduced the over-representation analysis of circRNAs via DisGeNET external biological database to find their potential molecular functions in neurodegenerative diseases [26]. CrowdGO provided an improvement in gene functional annotation with model-informed methods. Calculated GO term-semantic similarities are evaluated with a machine learning model to enhance the performance of consensus results [27]. Another study performed by Kumar et al. combined GO and KEGG terms for comprehensive enrichment analysis and visualized them with network topology-based approaches [28]. Contrary to the single knowledge base approach, Perscheid et al. introduced a novel method that integrates knowledge from curated databases and conventional gene selection approaches. The presented framework has achieved better classification accuracy [29].

Yousef and others recently introduced machine learning approaches based on grouping, scoring, and modeling (G-S-M) for gene expression analysis with biological information. They proposed various tools that follow this approach. For instance, maTE [30] adopts a biological grouping approach via integrating microRNAs (miRNAs). The GEO datasets and miRTarBase are given as input and RF model is trained with group information to model miRNA and mRNA regulations. The cogNet [31] serves as ranking active subnetworks and suggesting significant pathways by using KEGG pathways biological information. Another proposed tool, miRcorrNet [32], identifies miRNA-mRNAs regulatory modules via correlation analysis of expression profiles. The miRNA and mRNA profiles of target disease are retrieved from TCGA and fully data driven biological domain analysis is performed via G-S-M approach. The tool miRModuleNet [33] similar to miRcorrNet also detects significant miRNA-mRNA groups by considering two omics datasets. The relationships of pairs are calculated by Mutual Information which differs from the previous tool using correlation function. The significant groups ranking is not only based on the gene list but also miRNA information. Another G-S-M model-based study by Yousef et al. [34] integrates Gene Ontology information for grouping the genes. A novel approach PriPath [35] utilizes ranking and grouping functions to analyze gene expression with KEGG pathways. GediNET [36] incorporates gene information associated with diseases like cancer to identify significant groups. For identification of disease-disease associations, “disease is represented by a list of genes” strategy is used. The RF classifier is trained, and performance results are evaluated with Area Under Curve (AUC). The approaches like GediNET enable the improvement of disease diagnosis, prognosis, and treatment.

The idea of considering groups or clusters of genes instead of individual genes in studies was pioneered by Yousef et al., followed by more studies to improve the tools [37], [38]. Similarly, Support Vector Machine with Recursive Network Elimination (SVM-RNE) [39] method integrates gene network information by using the G-S-M model. Table 1 gives summaries of the main tools with type of the method, disease in case study, and biological knowledge details in this review.

Integrative approaches for multi-omics data

Understanding the functioning of biological systems with heterogeneous characteristics has directed scientists to deeper analyses of omics data. As illustrated in Figure 1, a wealth of data repositories providing valuable building blocks and biological samples take integration ap-

Table 1: The summaries of the main tools including type of the method, disease in case study, and biological knowledge

Tool/ Study name	Method Machine Learning (ML), Statistical Analysis (SA)	Disease in case study	Biological Knowledge: Pre-existing (PE), Fully data driven (FDD)	Link	Information
DiVenn	SA: Directed graph (modified Fisher Exact test)	bacterial pathogen on Arabidopsis plants	PE + FDD: GO and KEGG, RNA-Seq	http://divenn.noble.org/	Focus on the gene regulation levels for each gene and integrates KEGG pathway and gene ontology knowledge for the data visualization
Mergeomics 2.0	SA: Weighted correlation network analysis	20 Diseases (from metabolic syndrome to psychiatric disorders)	PE + FDD: GWAS datasets, Pathway databases (KEGG, Reactome, Biocarta, etc.)	http://mergeomics.research.idre.ucla.edu/	Multi-omics data integration to elucidate disease networks and predict therapeutics
Wang et al. [26]	SA: Weighted correlation network analysis	Fetal hippocampus of Down Syndrome Patients	PE + FDD: GO, ceRNAs and scRNA	–	Study the expression profiles and potential function of circRNAs in the fetal hippocampus from DS patients
CogNet	ML: RF	10 Disease (Cancer, Hypertension, Celiac, etc.)	PE + FDD: KEGG, GEO	https://github.com/malikyousef/miRcorrNet	Classify gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis
CrowdGO	ML: Adaptive Boosting	12 model and non-model species	PE + FDD: UniProt, Proteome	https://gitlab.com/mreijnders/CrowdGO	Consensus-based Gene Ontology (GO) term meta-predictor
DeepProg	ML: SVM and Deep Learning (GaussianMixture)	32 Cancer Diseases (TCGA)	FDD: miRNA-Seq, RNA-Seq and protein expression	https://github.com/lanagarmire/DeepProg	Predict patient survival subtypes using multi-omics data
mATE	ML: SVM-RCE	10 Disease (Cancer, Hypertension, Celiac, etc.)	PE + FDD: miRTarBase, GEO	https://github.com/malikyousef/maTE	Discover expressed interactions between microRNAs and their targets
mixOmics	ML: sPLS-DA (for DIABLO and MINT methods)	BRCA, transcriptomics stem cell studies	FDD: miRNA-Seq, RNA-Seq and protein expression	http://mixomics.org/	Multivariate projection-based methodologies
miRModuleNet	ML: RF	Cancer (TCGA)	FDD: miRNA-Seq and RNA-Seq	https://github.com/malikyousef/miRModuleNet/	Provide a hierarchical list of significant miRNA-mRNA regulatory modules
MiBiOmics	ML: sPLS-DA	Breast Cancer (TCGA)	FDD: miRNA-Seq, RNA-Seq and protein expression	https://gitlab.univ-nantes.fr/combi-ls2n/mibiomics	Web application for multi-omics data exploration and integration
Perscheid et al. [29]	ML: SVM-RFE, Variance-based, Information Gain, ReliefF	Cancer (TCGA)	PE + FDD: DisGeNET and KEGG, RNA-Seq	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6798862/	Automatic external knowledge integration, gene selection, and evaluation
STATegRa	ML: Component Analysis	Glioblastoma and the Skin Cutaneous Melanoma (TCGA)	FDD: miRNA-Seq, RNA-Seq and DNA methylation	https://bioconductor.org/packages/release/bioc/html/STATegRa.html	Integrative multi-omics analysis framework for identifying features and components

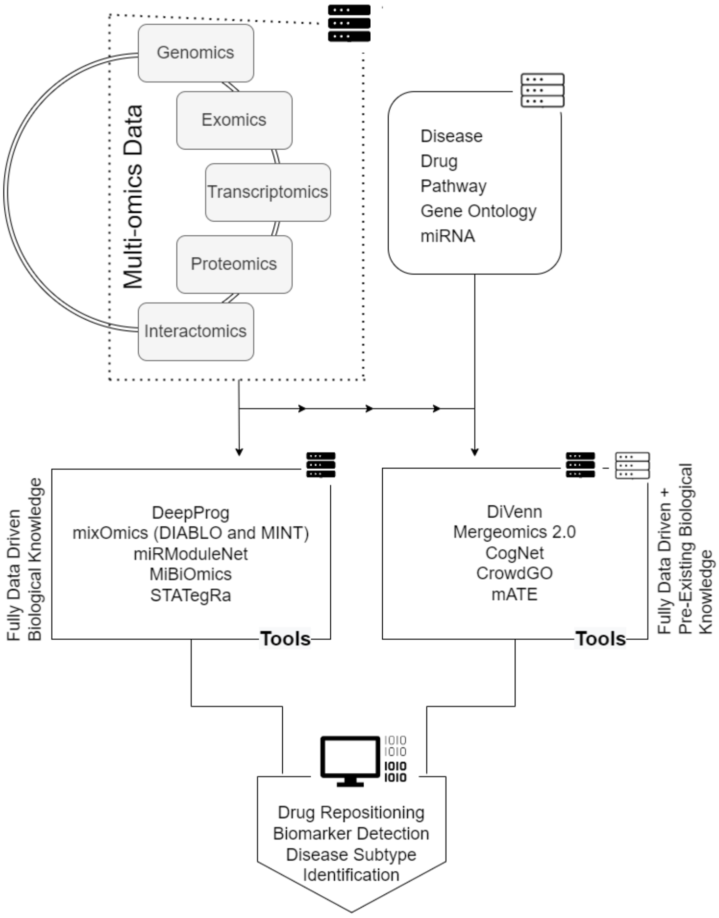

Figure 1: General Multi-omics data analysis framework, integrating omics data and pre-existing biological knowledge

proaches a step forward. Tools that adopt omics data features such as genomics, epigenomics, and metabolomics are required for the interpretation of affecting mechanisms of diseases in terms of genetic mutations, metabolites, and pathways etc. Advanced tools provisioning multi-omics data analysis can enable users to capture possible key factors associated with the phenotype of interest [40].

Deciphering these markers and their interplay can help to dissect the mechanism underlying disease onset and progression. Recently, proposed tools integrating multi-omics data are basically categorized as Bayesian, network, similarity, multivariate, supervised, semi-supervised, or unsupervised based approaches [41].

One of these tools, MiBiOmics, enables users to identify associations between up to 3 omics datasets. Network-based approach depending on weighted gene correlation network analysis is performed to explore molecular signatures and associations across layers [42]. STATegRa tools developed by Planell et al. combined feature identification with an unsupervised machine learning approach and detected enriched pathways with exploratory analysis [43]. The designed tool combines Principal Component Analysis, non-parametric combination for linking the features of different omics data with exploratory analysis. Mergeomics 2.0 presented by Ding et al. incorporates Meta marker set enrichment analysis for detection of omics-related disease pathways and networks through the integration of selected biomarkers. Subnetworks including gene sets associated with the interested disease are captured with key driver function and fed to PharmOmics repository for drug repositioning analysis [44].

mixOmics, a versatile multivariate method, enables the analysis of single and integrative omics data with modeling features as a set approach. The tool supports preprocessed multi-omics data from different platforms. The multivariate method is applied for the identification of molecular signatures and the distinction of disease subtypes via un/supervised analysis [45]. The frameworks DIABLO and MINT are developed for integration datasets. While DIABLO enables integration of same samples from different omics platforms, MINT integrates independent datasets.

Another machine learning tool, miRcorrNet, developed by Yousef et al. integrates miRNAs and gene expression profiles via a supervised machine learning approach. Highly scored groups, including target gene lists constructed with grouping functions, are utilized for the identification of disease-related biosignatures [32]. The following tool, miRModuleNet, integrates a pair of omics data to get more insight into the disease process. Generated hierarchical group list, each of the groups including miRNA and associated genes, with Mutual Information is introduced into machine learning model and intergroup relationships of the groups evaluated for deciphering significant therapeutic targets affecting disease progression [33].

DeepProg, semi-supervised hybrid ML tool, models patient survival to predict new patient statuses by combining deep learning and ML approaches. Multi-omics data matrices and survival information is given as input and cluster labels obtained by GaussianMixture function are used to build models via SVM to predict the subtypes of target disease [46].

Conclusion

In this review, we have surveyed several computational tools that tackle the integration of biological domain knowledge into the machine learning algorithm while in the second part the multi-omics computational tools were surveyed to open up new prospects for readers in the field. Multiple layer analysis of biological information leads to deeper understanding of biological systems. Strategies regarding the combination of fully data-driven and pre-existing biological knowledge in selecting features can improve the classification performance and potential marker selection. The tools using pre-existing knowledge in multi-omics integration may pave the way for a better comprehension in complex biological systems. Thus, extracting the biological knowledge from multi-omics datasets can be utilized to develop a novel integrative tool addressing multi-omics applications and study complex biological processes holistically.

Notes

Competing interests

The authors declare that they have no competing interests.

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nat Rev Methods Primer. 2021;1(59):1-21. DOI: 10.1038/s43586-021-00056-9
2. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan;30(1):207-10. DOI: 10.1093/nar/30.1.207
3. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. DOI: 10.1093/nar/gkq1019
4. Vizcaíno JA, Côté R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L. A guide to the Proteomics Identifications Database proteomics data repository. Proteomics. 2009 Sep;9(18):4276-83. DOI: 10.1002/pmic.200900402
5. Samaras P, Schmidt T, Frejno M, Gessulat S, Reinecke M, Jarzab A, Zecha J, Mergner J, Giansanti P, Ehrlich HC, Aiche S, Rank J, Kienegger H, Krcmar H, Kuster B, Wilhelm M. ProteomicsDB: a multi-omics and multi-organism resource for life science research. Nucleic Acids Res. 2020 Jan;48(D1):D1153-D1163. DOI: 10.1093/nar/gkz974
6. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database ConsortiumNeale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434-43. DOI: 10.1038/s41586-020-2308-7
7. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct – open source resource for molecular interaction data. Nucleic Acids Res. 2007 Jan;35(Database issue):D561-5. DOI: 10.1093/nar/gkl958
8. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, Dolma S, Coulombe-Huntington J, Chatr-Aryamontri A, Dolinski K, Tyers M. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021 Jan;30(1):187-200. DOI: 10.1002/pro.3978
9. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021 Jan;49(D1):D605-D612. DOI: 10.1093/nar/gkaa1074
10. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014 Mar;11(3):333-7. DOI: 10.1038/nmeth.2810
11. Koh HWL, Fermin D, Vogel C, Choi KP, Ewing RM, Choi H. iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. NPJ Syst Biol Appl. 2019;5:22. DOI: 10.1038/s41540-019-0099-y
12. Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10:2. DOI: 10.1186/s13040-017-0126-8
13. Tjärnberg A, Mahmood O, Jackson CA, Saldi GA, Cho K, Christiaen LA, Bonneau RA. Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data. PLoS Comput Biol. 2021 Jan;17(1):e1008569. DOI: 10.1371/journal.pcbi.1008569
14. Acharjee A, Larkman J, Xu Y, Cardoso VR, Gkoutos GV. A random forest based biomarker discovery and power analysis framework for diagnostics research. BMC Med Genomics. 2020 Nov;13(1):178. DOI: 10.1186/s12920-020-00826-6
15. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Sauter G. Gene-expression profiles in hereditary breast cancer. N Engl J Med. 2001 Feb;344(8):539-48. DOI: 10.1056/NEJM200102223440801

16. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 2000 Aug;406(6795):536-40. DOI: 10.1038/35020115

17. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol. 1999;6(3-4):281-97. DOI: 10.1089/106652799318274

18. Yousef M, Najami N, Abedallah L, Khalifa W. Computational Approaches for Biomarker Discovery. J Intell Learn Syst Appl. 2014;6(4):153-61. DOI: 10.4236/jilsa.2014.64012

19. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, Tsai TR, Huang HT, Wang CY, Wu HY, Ho SY, Chen PR, Chuang CH, Hsieh PJ, Wu YS, Chen WL, Li MJ, Wu YC, Huang XY, Ng FL, Buddhakosai W, Huang PC, Lan KC, Huang CY, Weng SL, Cheng YN, Liang C, Hsu WL, Huang HD. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2018 Jan;46(D1):D296-D302. DOI: 10.1093/nar/gkx1067

20. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan;47(D1):D330-D338. DOI: 10.1093/nar/gky1055

21. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan;45(D1):D353-D361. DOI: 10.1093/nar/gkw1092

22. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020 Jan;48(D1):D845-D855. DOI: 10.1093/nar/gkz1021

23. Bellazzi R, Zupan B. Towards knowledge-based gene expression data mining. J Biomed Inform. 2007 Dec;40(6):787-802. DOI: 10.1016/j.jbi.2007.06.005

24. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics. 2014;15 (Suppl 2):S2. DOI: 10.1186/1471-2105-15-S2-S2

25. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007 Jan;23(2):257-8. DOI: 10.1093/bioinformatics/btl567

26. Wang S, Tang X, Qin L, Shi W, Bian S, Wang Z, Wang Q, Wang X, Gu J, Hao B, Ding K, Liao S. Integrative Analysis Extracts a Core ceRNA Network of the Fetal Hippocampus With Down Syndrome. Front Genet. 2020;11:565955. DOI: 10.3389/fgene.2020.565955

27. Reijnders MJMF, Waterhouse RM. CrowdGO: Machine learning and semantic similarity guided consensus Gene Ontology annotation. PLoS Comput Biol. 2022 May;18(5):e1010075. DOI: 10.1371/journal.pcbi.1010075

28. Udhaya Kumar S, Thirumal Kumar D, Bithia R, Sankar S, Magesh R, Sidenna M, George Priya Doss C, Zayed H. Analysis of Differentially Expressed Genes and Molecular Pathways in Familial Hypercholesterolemia Involved in Atherosclerosis: A Systematic and Bioinformatics Approach. Front Genet. 2020;11:734. DOI: 10.3389/fgene.2020.00734

29. Perscheid C, Grasnick B, Uflacker M. Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches. J Integr Bioinform. 2018 Dec;16(1). DOI: 10.1515/jib-2018-0064

30. Yousef M, Abdallah L, Allmer J. maTE: discovering expressed interactions between microRNAs and their targets. Bioinformatics. 2019 Oct;35(20):4020-8. DOI: 10.1093/bioinformatics/btz204

31. Yousef M, Ülgen E, Uğur Sezerman O. CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. PeerJ Comput Sci. 2021;7:e336. DOI: 10.7717/peerj-cs.336

32. Yousef M, Goy G, Mitra R, Eischen CM, Jabeer A, Bakir-Gungor B. miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking. PeerJ. 2021;9:e11458. DOI: 10.7717/peerj.11458

33. Yousef M, Goy G, Bakir-Gungor B. miRModuleNet: Detecting miRNA-mRNA Regulatory Modules. Front Genet. 2022;13:767455. DOI: 10.3389/fgene.2022.767455

34. Yousef M, Sayıcı A, Bakir-Gungor B. Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In: Kotsis G, Tjoa AM, Khalil I, Moser B, Mashkoor A, Sametinger J, Fensel A, Martinez-Gil J, Fischer L, Czech G, Sobieczky F, Khan S, editors. Database and Expert Systems Applications – DEXA 2021 Workshops. Cham: Springer International Publishing; 2021. (Communications in Computer and Information Science; 1479). p. 205-14. DOI: 10.1007/978-3-030-87101-7_20

35. Yousef M, Ozdemir F, Jaaber A, Allmer J, Bakir-Gungor B. PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach [Preprint]. Research Square. 2022 Apr. DOI: 10.21203/rs.3.rs-1449467/v1

36. Yousef M, Qumsiyeh E. GediNET – Discover Disease-Disease Gene Associations utilizing Knowledge-based Machine Learning [Preprint]. Research Square. 2022 May. DOI: 10.21203/rs.3.rs-1643219/v1

37. Yousef M, Jung S, Showe LC, Showe MK. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. BMC Bioinformatics. 2007 May;8:144. DOI: 10.1186/1471-2105-8-144

38. Yousef M, Bakir-Gungor B, Jabeer A, Goy G, Qureshi R, C Showe L. Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME. F1000Res. 2020;9:1255. DOI: 10.12688/f1000research.26880.2

39. Yousef M, Ketany M, Manevitz L, Showe LC, Showe MK. Classification and biomarker identification using gene network modules and support vector machines. BMC Bioinformatics. 2009 Oct;10:337. DOI: 10.1186/1471-2105-10-337

40. Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS 2nd, Byrum SD. Multi-omics data integration considerations and study design for biological systems and disease. Mol Omics. 2021 Apr;17(2):170-85. DOI: 10.1039/d0mo00041h

41. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:1177932219899051. DOI: 10.1177/1177932219899051

42. Zoppi J, Guillaume JF, Neunlist M, Chaffron S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. BMC Bioinformatics. 2021 Jan;22(1):6. DOI: 10.1186/s12859-020-03921-8

43. Planell N, Lagani V, Sebastian-Leon P, van der Kloet F, Ewing E, Karathanasis N, Urdangarin A, Arozarena I, Jagodic M, Tsamardinos I, Tarazona S, Conesa A, Tegner J, Gomez-Cabrero D. STATegra: Multi-Omics Data Integration - A Conceptual Scheme With a Bioinformatics Pipeline. Front Genet. 2021;12:620453. DOI: 10.3389/fgene.2021.620453

44. Ding J, Blencowe M, Nghiem T, Ha SM, Chen YW, Li G, Yang X. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Res. 2021 Jul;49(W1):W375-W387. DOI: 10.1093/nar/gkab405

45. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ‘omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov;13(11):e1005752. DOI: 10.1371/journal.pcbi.1005752

46. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med. 2021 Jul;13(1):112. DOI: 10.1186/s13073-021-00930-x

Corresponding author:

Malik Yousef
Zefat Academic College, Jerusalem St 11, Zefat, 1320611, Israel
malik.yousef@gmail.com

Please cite as

Unlu Yazici M, Bakir-Gungor B, Yousef M. Integrative analyses in omics data: Machine learning perspective. GMS Med Inform Biom Epidemiol. 2023;19:Doc05.
DOI: 10.3205/mibe000244, URN: urn:nbn:de:0183-mibe0002442

This article is freely available from

https://doi.org/10.3205/mibe000244

Published: *2023-07-04*