Report Title

Introduction

Introduction:
The early detection of complex diseases, such as cancer, has been a longstanding challenge in biomedical research. The utilization of integrative genomics and multi-omics data analysis has emerged as a promising approach to identify novel biomarkers for early disease detection. This research aims to address the pressing need for more effective methods to detect complex diseases at an early stage, thereby revolutionizing early detection and treatment strategies for these diseases.

Current studies have focused on leveraging various molecular data types, such as gene expression, miRNA expression, DNA methylation, and CNV data, to predict clinical target variables and classify cancer subtypes. The significance of this research lies in its potential to improve early detection and prognosis of cancer, ultimately leading to more personalized and effective treatments. The comprehensive evaluation of 16 deep learning methods on simulated, single-cell, and cancer multi-omics datasets emphasizes the importance of this research. Benchmarking results have highlighted the performance of methods such as moGAT in achieving the best classification performance and efmmdVAE, efVAE, and lfmmdVAE in clustering tasks, demonstrating the potential impact of integrative genomics and multi-omics data analysis on clinical practice and patient outcomes.

The study not only provides valuable insights for biomedical researchers in selecting appropriate multi-omics data fusion methods but also paves the way for future advancements in the field. Furthermore, the research addresses the critical need for accurate and reliable early detection biomarkers through the integration of genomics and multi-omics data analysis, potentially impacting the development of precision medicine and improved patient outcomes. This study's findings can serve as a reference for biomedical researchers to choose appropriate deep learning-based multi-omics data fusion methods and suggests future directions in this field, emphasizing the critical role of integrative genomics and multi-omics data analysis in advancing early detection strategies for complex diseases, particularly cancer.

References:

1. Source URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9907220/pdf/
2. Source URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10484010/pdf/
3. Source URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9929204/pdf/
4. Source URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9361561/pdf/
5. Source URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8187432/pdf/

Literature

(TCGA) cancer multi-omics datasets, which consist of three omics data types. Evaluation metrics such as accuracy, F1 macro, F1 weighted, Jaccard Index, C-index, and Silhouette score are used for classification and clustering tasks.

The experimental results indicate promising performance of DL-based multi-omics fusion methods in classifying samples with different cancer subtypes. GNN-based methods also show effectiveness in classifying different cancer subtypes, with moGAT performing well in breast cancer (BRCA) and glioblastoma (GBM), while moGCN, lfNN, and efNN excel in sarcoma (SARC), lung adenocarcinoma (LUAD), and stomach cancer (STAD) respectively.

Furthermore, this research demonstrates the potential of DL-based multi-omics fusion methods in predicting clinical target variables in various cancer types. The integration of integrative genomics and multi-omics data analysis with DL-based methods has shown great potential in identifying novel biomarkers for early detection of complex diseases, particularly cancer. These innovative approaches could potentially pave the way for more precise and personalized diagnostic and therapeutic strategies.

In conclusion, the literature review highlights the experimental methods, settings, and results of the application of integrative genomics and multi-omics data analysis for the early detection of complex diseases, such as cancer. The article showcases the potential of DL-based multi-omics fusion methods to identify novel biomarkers, providing insight into the molecular and clinical characteristics of these diseases.

Discussion

Discussion:

The research articles discussed focus on the utilization of integrative genomics and multi-omics data analysis to identify novel biomarkers for early detection of complex diseases, particularly cancer. Various studies explore the application of Graph Neural Networks (GNN)-based methods, such as moGAT, moGCN, and MOGONET, in cancer classification tasks using multi-omics datasets.

Strengths Identified Across the Studies:

- Utilization of GNN-based methods: The studies leverage advanced GNN-based methods like moGAT, moGCN, and MOGONET to effectively analyze multi-omics datasets, showcasing the potential for these techniques in disease classification and biomarker identification.
- Comprehensive evaluation metrics: There is a common theme of employing a diverse set of evaluation metrics, including accuracy, F1 scores, Jaccard Index, and more, demonstrating a thorough assessment of the proposed methods' performance in classification and clustering tasks.
- Superior performance in biomarker identification: The studies highlight the ability of GNN-based methods and MOGONET to outperform existing approaches in identifying important omics signatures and biomarkers associated with various biomedical problems, showcasing their potential for early disease detection.

Weaknesses Identified in the Studies:

- Limited real-world validation: The reliance on simulated datasets in some studies may restrict the generalizability of the findings to real-world clinical applications, indicating the need for validation using real-world clinical datasets.
- Lack of comparison with existing methods: Although the studies demonstrate the superiority of GNN-based methods, a direct comparison with other state-of-the-art approaches in the field may provide a more comprehensive understanding of the proposed methods' effectiveness.
- Inadequate clinical relevance: Some studies do not explicitly discuss the clinical implications of the identified biomarkers or their impact on patient outcomes, highlighting the need for further exploration of the clinical relevance of the findings.

Future Research Directions:

- Validation on real-world clinical datasets: Future research should focus on validating the proposed methods using real-world clinical datasets to ensure their applicability in clinical settings and enhance their relevance.
- Comparative studies: Conducting comparative studies with a wider range of existing methods would facilitate a more in-depth evaluation of the strengths and limitations of GNN-based approaches and MOGONET in multi-omics data analysis.
- Integration of clinical outcomes: Future studies could integrate clinical outcome data to assess the identified biomarkers' impact on disease prognosis and treatment response, contributing to a comprehensive understanding of their clinical relevance and utility.

In conclusion, the studies contribute significantly to the field of integrative genomics and multi-omics data analysis by introducing innovative methods for identifying biomarkers for early disease detection. While showcasing promising results in biomarker identification, the studies also underscore the importance of addressing limitations through real-world validation, comparative analyses, and exploration of clinical implications. Continued research efforts are warranted to advance the application of multi-omics data analysis in improving early disease detection and patient outcomes.

Idea

Problem:
Exploring the development of a comprehensive multi-omics data fusion framework that leverages deep learning methodologies to integrate heterogeneous biomedical data, with a focus on addressing the challenges of data sparsity, multimodal interpretability, and standardization, ultimately aiming to enhance biomarker discovery and patient classification in precision medicine.

Rationale:

The emergence of high-dimensional, high-throughput multi-scale biomedical data necessitates the development of advanced computational methods for integrating diverse omics data types. By leveraging deep learning approaches, this research problem seeks to address the persistent challenges of data sparsity, multimodal interpretability, and standardization while aiming to enhance the identification of robust biomarkers and improve patient classification in the context of precision

medicine. This problem aligns with the growing importance of personalized medicine and the need for comprehensive, innovative data fusion frameworks to fully harness the potential of multi-omics data in biomedical applications.

Method

Method: Multi-scale Deep Fusion Framework (M-DFF)

Rationale:

The Multi-scale Deep Fusion Framework (M-DFF) addresses the research problem by integrating heterogeneous biomedical data through a comprehensive and innovative approach that leverages deep learning methodologies. M-DFF emphasizes the need to overcome the challenges of data sparsity, multimodal interpretability, and standardization while aiming to enhance biomarker discovery and patient classification in precision medicine. The rationale behind M-DFF is deeply rooted in the emergence of high-dimensional, high-throughput multi-scale biomedical data, necessitating the development of advanced computational methods for integrating diverse omics data types, in alignment with the growing importance of personalized medicine.

Method Description:

1. Data Preprocessing and Feature Extraction: M-DFF begins with the preprocessing of multi-omics data, including mRNA expression data, DNA methylation data, and microRNA expression data. Feature extraction techniques such as dimensionality reduction and normalization are applied to address data sparsity and standardization challenges.

2. Multi-modal Data Integration: M-DFF implements a multi-modal data integration strategy that combines omics-specific learning and cross-omics correlation learning to effectively fuse the diverse data types, enabling comprehensive analysis across multiple scales.

3. Graph-based Representation: A graph-based representation of the integrated multi-omics data is constructed, utilizing graph convolutional networks (GCNs) to capture complex interdependencies and interactions between heterogeneous data elements.

4. Deep Learning Model Integration: The framework incorporates deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to leverage the multi-scale nature of the data and facilitate interpretability across modalities.

5. Biomarker Discovery and Patient Classification: M-DFF employs the integrated deep learning models to identify robust biomarkers and improve patient classification in the context of precision medicine, thereby enabling the translation of multi-omics data into actionable clinical insights.

6. Evaluation and Validation: Comprehensive evaluation metrics, including accuracy, F1 score, and area under the curve (AUC), are utilized to rigorously assess the performance of M-DFF in biomarker discovery and patient classification tasks. Moreover, external validation using independent datasets ensures the generalizability of the framework.

In summary, the Multi-scale Deep Fusion Framework (M-DFF) represents an innovative and rigorous methodology for the comprehensive integration and analysis of multi-omics data, leveraging deep learning approaches to address the challenges of data sparsity, multimodal interpretability, and standardization. M-DFF offers a generalizable solution that has the potential to advance biomarker discovery and patient classification in the realm of precision medicine, thereby contributing to the realization of personalized and targeted therapeutic interventions.

Experiment

Experiment:

Rationale:

The multi-scale deep fusion framework (M-DFF) proposed to address the challenges of data sparsity, multimodal interpretability, and standardization in the integration of heterogeneous biomedical data is a promising approach in the field of precision medicine. To validate the effectiveness and generalizability of the M-DFF method for biomarker discovery and patient classification tasks, an experiment combining data simulation, model training, and performance evaluation across various datasets is designed.

Experimental Design:

1. **Data Simulation and Preprocessing:**

- Simulate multi-omics datasets incorporating mRNA expression, DNA methylation, and microRNA data with varying levels of data sparsity, noise, and inter-omics correlations to mimic real-world scenarios.
- Perform data preprocessing steps including normalization, feature extraction, and dimensionality reduction as outlined in the M-DFF method description.

2. **Model Training and Validation:**

- Implement the M-DFF methodology for multi-modal data integration using graph-based representation and deep learning model integration on the simulated datasets.
- Train deep learning models such as CNNs and RNNs on the integrated data to identify biomarkers and classify patients in the context of precision medicine.
- Utilize evaluation metrics such as accuracy, F1 score, and AUC to assess the performance of the

M-DFF framework in biomarker discovery and patient classification tasks.

3. **Cross-Dataset Validation:**

- Validate the trained M-DFF model on independent datasets sourced from related studies to ensure the generalizability of the framework.
- Compare the performance of M-DFF with other state-of-the-art methods as mentioned in the related papers for benchmarking purposes.

4. **Sensitivity Analysis and Robustness Testing:**

- Conduct sensitivity analysis by varying key parameters within the M-DFF framework to evaluate its robustness and performance under different settings.
- Test the robustness of the model against data perturbations, missing values, and variations in omics data distributions.

5. **Interpretability Assessment:**

- Explore the interpretability of the deep learning models integrated within M-DFF using visualization techniques such as saliency maps, feature importance plots, and attention mechanisms to understand the biomarker identification process.

Expected Outcome:

The experiment aims to validate the efficacy of the M-DFF method in addressing the challenges of multi-omics data integration for biomarker discovery and patient classification in precision medicine. By systematically evaluating the framework's performance across diverse datasets and conducting sensitivity analyses, this study can provide insights into the strengths and limitations of the M-DFF approach, paving the way for its potential application in real-world biomedical scenarios.

More related paper

Paper 1
Title: Dealing with dimensionality: the application of machine learning to multi-omics data.

Abstract: MOTIVATION: Machine learning (ML) methods are motivated by the need to automate information extraction from large datasets in order to support human users in data-driven tasks. This is an attractive approach for integrative joint analysis of vast amounts of omics data produced in next generation sequencing and other -omics assays. A systematic assessment of the current literature can help to identify key trends and potential gaps in methodology and applications. We surveyed the literature on ML multi-omic data integration and quantitatively explored the goals, techniques and data involved in this field. We were particularly interested in examining how researchers use ML to deal with the volume and complexity of these datasets. RESULTS: Our main finding is that the methods used are those that address the challenges of datasets with few samples and many features. Dimensionality

reduction methods are used to reduce the feature count alongside models that can also appropriately handle relatively few samples. Popular techniques include autoencoders, random forests and support vector machines. We also found that the field is heavily influenced by the use of The Cancer Genome Atlas dataset, which is accessible and contains many diverse experiments. AVAILABILITY AND IMPLEMENTATION: All data and processing scripts are available at this GitLab repository: https://gitlab.com/polavieja_lab/ml_multi-omics_review/ or in Zenodo: https://doi.org/10.5281/zenodo.7361807. SUPPLEMENTARY INFORMATION: Supplementary data are available at Bioinformatics online.

DOI: 10.1093/bioinformatics/btad021

The impact factor: 6.931

Paper 2

Title: Multimodal data fusion for cancer biomarker discovery with deep learning.

Abstract: Technological advances now make it possible to study a patient from multiple angles with high-dimensional, high-throughput multi-scale biomedical data. In oncology, massive amounts of data are being generated ranging from molecular, histopathology, radiology to clinical records. The introduction of deep learning has significantly advanced the analysis of biomedical data. However, most approaches focus on single data modalities leading to slow progress in methods to integrate complementary data types. Development of effective multimodal fusion approaches is becoming increasingly important as a single modality might not be consistent and sufficient to capture the heterogeneity of complex diseases to tailor medical care and improve personalised medicine. Many initiatives now focus on integrating these disparate modalities to unravel the biological processes involved in multifactorial diseases such as cancer. However, many obstacles remain, including lack of usable data as well as methods for clinical validation and interpretation. Here, we cover these current challenges and reflect on opportunities through deep learning to tackle data sparsity and scarcity, multimodal interpretability, and standardisation of datasets.

DOI: 10.1038/s42256-023-00633-5

The impact factor: 25.898

Paper 3

Title: Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine.

Abstract: Cancer progression is linked to gene-environment interactions that alter cellular homeostasis. The use of biomarkers as early indicators of disease manifestation and progression can substantially improve diagnosis and treatment. Large omics datasets generated by high-throughput profiling technologies, such as microarrays, RNA sequencing, whole-genome shotgun sequencing, nuclear magnetic resonance, and mass spectrometry, have enabled data-driven biomarker discoveries. The

identification of differentially expressed traits as molecular markers has traditionally relied on statistical techniques that are often limited to linear parametric modeling. The heterogeneity, epigenetic changes, and high degree of polymorphism observed in oncogenes demand biomarker-assisted personalized medication schemes. Deep learning (DL), a major subunit of machine learning (ML), has been increasingly utilized in recent years to investigate various diseases. The combination of ML/DL approaches for performance optimization across multi-omics datasets produces robust ensemble-learning prediction models, which are becoming useful in precision medicine. This review focuses on the recent development of ML/DL methods to provide integrative solutions in discovering cancer-related biomarkers, and their utilization in precision medicine.

DOI: 10.1016/j.csbj.2023.01.043

The impact factor: 6.155

Paper 4

Title: A benchmark study of deep learning-based multi-omics data fusion methods for cancer.

Abstract: BACKGROUND: A fused method using a combination of multi-omics data enables a comprehensive study of complex biological processes and highlights the interrelationship of relevant biomolecules and their functions. Driven by high-throughput sequencing technologies, several promising deep learning methods have been proposed for fusing multi-omics data generated from a large number of samples. RESULTS: In this study, 16 representative deep learning methods are comprehensively evaluated on simulated, single-cell, and cancer multi-omics datasets. For each of the datasets, two tasks are designed: classification and clustering. The classification performance is evaluated by using three benchmarking metrics including accuracy, F1 macro, and F1 weighted. Meanwhile, the clustering performance is evaluated by using four benchmarking metrics including the Jaccard index (JI), C-index, silhouette score, and Davies Bouldin score. For the cancer multi-omics datasets, the methods' strength in capturing the association of multi-omics dimensionality reduction results with survival and clinical annotations is further evaluated. The benchmarking results indicate that moGAT achieves the best classification performance. Meanwhile, efmmdVAE, efVAE, and lfmmdVAE show the most promising performance across all complementary contexts in clustering tasks. CONCLUSIONS: Our benchmarking results not only provide a reference for biomedical researchers to choose appropriate deep learning-based multi-omics data fusion methods, but also suggest the future directions for the development of more effective multi-omics data fusion methods. The deep learning frameworks are available at https://github.com/zhenglinyi/DL-mo .

DOI: 10.1186/s13059-022-02739-2

The impact factor: 17.906

Paper 5

Title: Deep-Learning Algorithm and Concomitant Biomarker Identification for NSCLC Prediction Using Multi-Omics Data Integration.

Abstract: Early diagnosis of lung cancer to increase the survival rate, which is currently at a low range of mid-30%, remains a critical need. Despite this, multi-omics data have rarely been applied to non-small-cell lung cancer (NSCLC) diagnosis. We developed a multi-omics data-affinitive artificial intelligence algorithm based on the graph convolutional network that integrates mRNA expression, DNA methylation, and DNA sequencing data. This NSCLC prediction model achieved a 93.7% macro F1-score, indicating that values for false positives and negatives were substantially low, which is desirable for accurate classification. Gene ontology enrichment and pathway analysis of features revealed that two major subtypes of NSCLC, lung adenocarcinoma and lung squamous cell carcinoma, have both specific and common GO biological processes. Numerous biomarkers (i.e., microRNA, long non-coding RNA, differentially methylated regions) were newly identified, whereas some biomarkers were consistent with previous findings in NSCLC (e.g., SPRR1B). Thus, using multi-omics data integration, we developed a promising cancer prediction algorithm.

DOI: 10.3390/biom12121839

The impact factor: 6.064

Paper 6

Title: Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer.

Abstract: Cancer is defined as a large group of diseases that is associated with abnormal cell growth, uncontrollable cell division, and may tend to impinge on other tissues of the body by different mechanisms through metastasis. What makes cancer so important is that the cancer incidence rate is growing worldwide which can have major health, economic, and even social impacts on both patients and the governments. Thereby, the early cancer prognosis, diagnosis, and treatment can play a crucial role at the front line of combating cancer. The onset and progression of cancer can occur under the influence of complicated mechanisms and some alterations in the level of genome, proteome, transcriptome, metabolome etc. Consequently, the advent of omics science and its broad research branches (such as genomics, proteomics, transcriptomics, metabolomics, and so forth) as revolutionary biological approaches have opened new doors to the comprehensive perception of the cancer landscape. Due to the complexities of the formation and development of cancer, the study of mechanisms underlying cancer has gone beyond just one field of the omics arena. Therefore, making a connection between the resultant data from different branches of omics science and examining them in a multi-omics field can pave the way for facilitating the discovery of novel prognostic, diagnostic, and therapeutic approaches. As the volume and complexity of data from the omics studies in cancer are increasing dramatically, the use of leading-edge technologies such as machine learning can have a promising role in the assessments of cancer research resultant data. Machine learning is categorized as a subset of artificial intelligence which aims to data parsing, classification, and data pattern identification by applying statistical methods and algorithms. This acquired knowledge subsequently allows computers to learn and improve accurate predictions through experiences from data processing. In this context, the application of machine learning, as a novel computational technology offers new opportunities for achieving in-depth knowledge of cancer by analysis of resultant data from multi-omics studies. Therefore, it can be concluded that the use of artificial intelligence technologies such as machine learning can have revolutionary roles in the fight against cancer.

DOI: 10.3389/fgene.2022.824451

The impact factor: 4.772

Paper 7

Title: Integrative omics approaches for biosynthetic pathway discovery in plants.

Abstract: Covering: up to 2022With the emergence of large amounts of omics data, computational approaches for the identification of plant natural product biosynthetic pathways and their genetic regulation have become increasingly important. While genomes provide clues regarding functional associations between genes based on gene clustering, metabolome mining provides a foundational technology to chart natural product structural diversity in plants, and transcriptomics has been successfully used to identify new members of their biosynthetic pathways based on coexpression. Thus far, most approaches utilizing transcriptomics and metabolomics have been targeted towards specific pathways and use one type of omics data at a time. Recent technological advances now provide new opportunities for integration of multiple omics types and untargeted pathway discovery. Here, we review advances in plant biosynthetic pathway discovery using genomics, transcriptomics, and metabolomics, as well as recent efforts towards omics integration. We highlight how transcriptomics and metabolomics provide complementary information to link genes to metabolites, by associating temporal and spatial gene expression levels with metabolite abundance levels across samples, and by matching mass-spectral features to enzyme families. Furthermore, we suggest that elucidation of gene regulatory networks using time-series data may prove useful for efforts to unwire the complexities of biosynthetic pathway components based on regulatory interactions and events.

DOI: 10.1039/d2np00032f

The impact factor: 15.111

Paper 8

Title: A guide to multi-omics data collection and integration for translational medicine.

Abstract: The emerging high-throughput technologies have led to the shift in the design of translational medicine projects towards collecting multi-omics patient samples and, consequently, their integrated analysis. However, the complexity of integrating these datasets has triggered new questions regarding the appropriateness of the available computational methods. Currently, there is no clear consensus on the best combination of omics to include and the data integration methodologies required for their analysis. This article aims to guide the design of multi-omics studies in the field of translational medicine regarding the types of omics and the integration method to choose. We review articles that perform the integration of multiple omics measurements from patient samples. We identify five objectives in translational medicine applications: (i) detect disease-associated molecular patterns, (ii) subtype identification, (iii) diagnosis/prognosis, (iv) drug response prediction, and (v) understand regulatory processes. We describe common trends in the selection of omic types combined for different objectives and diseases. To guide the choice of data integration tools, we group them into the scientific objectives they aim to address. We describe the main computational methods adopted to achieve these objectives and present examples of tools. We compare tools based on how they deal with the

computational challenges of data integration and comment on how they perform against predefined objective-specific evaluation criteria. Finally, we discuss examples of tools for downstream analysis and further extraction of novel insights from multi-omics datasets.

DOI: 10.1016/j.csbj.2022.11.050

The impact factor: 6.155

Paper 9

Title: A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment.

Abstract: Accurate diagnosis is the key to providing prompt and explicit treatment and disease management. The recognized biological method for the molecular diagnosis of infectious pathogens is polymerase chain reaction (PCR). Recently, deep learning approaches are playing a vital role in accurately identifying disease-related genes for diagnosis, prognosis, and treatment. The models reduce the time and cost used by wet-lab experimental procedures. Consequently, sophisticated computational approaches have been developed to facilitate the detection of cancer, a leading cause of death globally, and other complex diseases. In this review, we systematically evaluate the recent trends in multi-omics data analysis based on deep learning techniques and their application in disease prediction. We highlight the current challenges in the field and discuss how advances in deep learning methods and their optimization for application is vital in overcoming them. Ultimately, this review promotes the development of novel deep-learning methodologies for data integration, which is essential for disease detection and treatment.

DOI: 10.3389/fgene.2023.1199087

The impact factor: 4.772

Paper 10

Title: Artificial intelligence in cancer target identification and drug discovery.

Abstract: Artificial intelligence is an advanced method to identify novel anticancer targets and discover novel drugs from biology networks because the networks can effectively preserve and quantify the interaction between components of cell systems underlying human diseases such as cancer. Here, we review and discuss how to employ artificial intelligence approaches to identify novel anticancer targets and discover drugs. First, we describe the scope of artificial intelligence biology analysis for novel anticancer target investigations. Second, we review and discuss the basic principles and theory of commonly used network-based and machine learning-based artificial intelligence algorithms. Finally, we showcase the applications of artificial intelligence approaches in cancer target identification and drug discovery. Taken together, the artificial intelligence models have provided us with a quantitative framework to study the relationship between network characteristics and cancer, thereby leading to the identification of potential anticancer targets and the discovery of novel drug candidates.

DOI: 10.1038/s41392-022-00994-0

The impact factor: 38.104