STM JOURNALS

Research & Reviews: Journal of
Computational Biology

https://journals.stmjournals.com/rrjocb/

ISSN: 2319-3433 (Online)
ISSN: 2349-3720 (Print)
Volume 13, Issue 3, 2024
September–December
DOI (Journal): 10.37591/RRJoCB

Review RRJoCB

Data Integration and Visualization in Bioinformatics: Techniques and Challenges

Mansi Srivastava*

Abstract

Data integration and visualization play essential roles in bioinformatics, facilitating the thorough analysis, and interpretation of intricate biological datasets. In the field of bioinformatics, vast amounts of data are generated from various experimental platforms, such as genomic sequencing, proteomics, transcriptomics, and metabolomics. However, the heterogeneity of these datasets, coupled with their large scale and complexity, presents significant challenges in terms of integration, analysis, and visualization. Data integration techniques aim to combine disparate datasets from multiple sources, enabling the creation of a unified view of biological systems. These methods include approaches, such as concatenation, matrix factorization, and multivariate statistics, along with more advanced techniques leveraging machine learning and deep learning algorithms. Integration is essential for revealing hidden relationships between various types of biological data and for generating holistic insights into the underlying biology. Visualization tools are crucial for transforming complex biological data into easily understandable formats. They transform raw data into intuitive visual representations, such as heatmaps, scatter plots, and network diagrams, which facilitate the identification of patterns, trends, and outliers. Advanced techniques like principal component analysis (PCA), t-SNE, and network-based visualizations are increasingly being used to represent multidimensional data. Despite the advancements, challenges persist in integrating and visualizing bioinformatics data. Issues, such as data heterogeneity, scalability, noise, and the curse of dimensionality complicate these tasks. Additionally, integrating multi-omics data and visualizing high-dimensional datasets continue to present major challenges. This article discusses the techniques used in data integration and visualization, highlights the challenges faced, and outlines the future directions in bioinformatics for overcoming these hurdles.

Keywords: Data integration, bioinformatics data, biological data, protein structures, genomic sequences

INTRODUCTION

Bioinformatics, an interdisciplinary field combining biology, computer science, and statistics, is essential to contemporary biological research. The primary objective of bioinformatics is to derive meaningful insights from large, complex biological datasets, frequently generated by high-throughput technologies like next-generation sequencing (NGS), mass spectrometry, and microarray analysis.

These datasets, encompassing genomic sequences, gene expression profiles, protein structures, and metabolite levels, provide rich information that enhances our understanding of biological processes and diseases. However, to fully realize their potential, these datasets must be integrated and analyzed in a comprehensive and

***Author for Correspondence**
Mansi Srivastava
E-mail: srivastavamansi2012@gmail.com

Student, Department of Biotechnology, Amity University Gurgaon, Haryana, India

Received Date: November 27, 2024
Accepted Date: December 02, 2024
Published Date: December 19, 2024

Citation: Mansi Srivastava. Data Integration and Visualization in Bioinformatics: Techniques and Challenges. Research & Reviews: Journal of Computational Biology. 2024; 13(3): 1–8p.

systematic manner. This is where data integration and visualization techniques come into play, providing powerful tools for processing, analyzing, and presenting biological data [1–3].

THE EXPLOSION OF BIOLOGICAL DATA

In recent years, the volume of biological data has expanded rapidly, fueled by advancements in high-throughput technologies. Genomic data, for instance, now includes billions of DNA sequences, producing vast datasets that were once inconceivable. Similarly, transcriptomic data, which measures gene expression levels, and proteomic data, which characterizes proteins and their interactions, have become increasingly complex and voluminous. Beyond genomics and proteomics, other data types, such as metabolomics, epigenomics, and microbiome data are also contributing to the growing complexity of biological research [4]. This rapid increase in data generation has presented both opportunities and challenges for the field of bioinformatics. While these rich datasets offer immense potential for advancing biological research and medicine, their integration poses significant challenges. Biological data is inherently heterogeneous, with various formats, scales, and levels of noise. Genomic data might come from sequencing machines that generate nucleotide sequences in a text-based format, whereas proteomic data may be represented in matrices containing protein abundance levels across various conditions. Transcriptomic data, typically measured using RNA sequencing, also comes with its own set of challenges, such as variation in sequencing depth, technical noise, and batch effects. Integrating these disparate data types into a unified, coherent framework is a critical step for extracting useful insights [5–9].

THE NEED FOR DATA INTEGRATION

Data integration in bioinformatics involves combining datasets from different sources and platforms to create a more comprehensive understanding of biological systems. This can take many forms, from integrating genomic data with clinical information to merging gene expression data with proteomic data for a deeper understanding of cellular functions. Integrating data from multiple sources is essential for several reasons [10].

First, biological systems are highly interconnected and understanding the relationships between different molecular levels – such as how gene expression affects protein synthesis, or how proteins influence metabolic pathways – is key to unraveling the complexity of biological phenomena. Second, integration can help reduce noise and improve the accuracy of predictions. For example, integrating genomic data with epigenomic and transcriptomic data can help validate genetic findings and uncover regulatory mechanisms that might not be evident from any single data type. Finally, integration is crucial for advancing personalized medicine, where data from multiple sources – such as patient genomics, clinical history, and environmental factors – can be used to develop tailored treatment strategies.

However, integrating multi-omics data (such as genomics, transcriptomics, and proteomics) demands advanced computational tools and techniques. Challenges, such as data heterogeneity, missing values, varying sample sizes, and differing experimental conditions complicate the integration process. Additionally, large-scale data integration often demands substantial computational resources, particularly when dealing with “big data” in bioinformatics [11–13].

THE ROLE OF DATA VISUALIZATION

Visualization is an essential tool for interpreting and presenting the results of integrated bioinformatics analyses. While computational techniques for data integration can help uncover complex biological relationships, visual representations are necessary to make these relationships comprehensible to researchers and clinicians. Visualization techniques help to reduce the complexity of high-dimensional biological data, revealing underlying patterns, correlations, and trends that might be difficult to discern otherwise.

In bioinformatics, the choice of visualization techniques varies depending on the data type and the specific goals of the research. For instance, heatmaps are often used to represent gene expression data

across different conditions, Scatter plots and volcano plots are commonly utilized to illustrate the relationship between gene expression levels and fold changes in differential analysis. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are widely used dimensionality reduction techniques that allow researchers to represent high-dimensional data in two or three dimensions for easier visualization. Network-based visualizations, such as protein-protein interaction (PPI) networks or gene co-expression networks, are frequently used to represent molecular relationships and biological pathways [14–19].

Interactive visualizations, enabling real-time data exploration and zooming into specific areas of interest, have gained significant popularity in bioinformatics. Tools, such as Cytoscape, D3.js, and Shiny enable the creation of dynamic, web-based visualizations that can be shared among researchers, facilitating collaboration, and data dissemination.

CHALLENGES IN DATA INTEGRATION AND VISUALIZATION

Despite significant advancements in bioinformatics tools and technologies, challenges remain in integrating and visualizing biological data, with the heterogeneity of such data being a major hurdle. Data from different experimental platforms and sources often differ in terms of format, scale, and quality, making integration difficult. Furthermore, biological data is inherently noisy, with missing values, measurement errors, and technical biases that can obscure meaningful patterns. The curse of dimensionality, where high-dimensional data becomes sparse and challenging to interpret, adds another layer of complexity to the analysis. Moreover, computational challenges, such as the need for scalable algorithms and efficient data storage, also hinder progress. The vast size of biological datasets, especially in the context of multi-omics data, requires innovative computational approaches to handle the data efficiently and ensure that integration and analysis processes are computationally feasible [19–23].

OBJECTIVE OF THE STUDY

This study aims to investigate the techniques and challenges involved in the integration and visualization of data in bioinformatics. With the growing complexity and volume of biological data generated through high-throughput technologies, this study aims to provide a comprehensive understanding of how diverse biological datasets can be integrated to extract meaningful insights. Additionally, the study seeks to highlight the role of data visualization in making these insights accessible and interpretable for researchers and clinicians [24–26].

The specific objectives of the study are as follows:

1. *To Examine the Techniques for Data Integration:* This study aims to explore the various approaches used for integrating biological data from different sources, including genomic, transcriptomic, proteomic, and metabolomic datasets. Key techniques, such as concatenation, matrix factorization, multivariate statistics, and advanced machine learning methods will be discussed. The goal is to provide an understanding of how these methods enable a holistic view of biological systems by combining data of varying formats and scales [27, 28].
2. *To Investigate the Role of Visualization in Bioinformatics:* The study seeks to evaluate the importance of data visualization in bioinformatics, particularly in the context of high-dimensional data. It will examine how visualization techniques, including heatmaps, scatter plots, PCA, and network-based visualizations, aid in the interpretation of complex biological datasets. The objective is to show how these visual tools help researchers identify trends, patterns, and outliers in multi-omics data [29, 30].
3. *To Identify and Analyze the Challenges in Data Integration and Visualization:* A major objective of this study is to investigate the challenges that arise during the integration and visualization of bioinformatics data. These challenges include issues, such as data heterogeneity, noise, missing values, scalability, and the curse of dimensionality. The study will aim to identify current gaps in methodologies and propose potential solutions to address these challenges [31, 32].

4. *To Highlight the Impact of Data Integration and Visualization in Biological Research:* The study intends to explore the significant role of data integration and visualization in advancing biological research, particularly in personalized medicine, systems biology, and drug discovery. By providing examples from real-world applications, the study will illustrate how integrated and visualized data can lead to new biological insights, improve understanding of disease mechanisms, and aid in clinical decision-making [33–37].
5. *To Explore Future Directions and Emerging Technologies:* Finally, the study will investigate emerging technologies and methodologies in bioinformatics that are enhancing data integration and visualization [38]. This includes the use of artificial intelligence (AI) and machine learning for better data integration, cloud computing for scalable analysis, and next-generation visualization tools that allow for more interactive, dynamic, and real-time data exploration. The objective is to provide an outlook on the future trends in bioinformatics that could overcome current limitations in data integration and visualization [39, 40].

Overall, the objective of this study is to provide a detailed overview of the state-of-the-art techniques in bioinformatics data integration and visualization, while also addressing the challenges and proposing solutions that will enable more effective and insightful analyses in the rapidly evolving field of biological research.

To provide a side-by-side comparison of popular data integration techniques, highlighting their strengths, applications, and challenges (Table 1).

Table 1. Comparison of bioinformatics data integration techniques.

Integration Technique	Description	Applications	Challenges
Concatenation	Merging datasets into a single dataset.	Merging genomic datasets from multiple samples.	Potential loss of context and complexity.
Matrix Factorization	Decomposing data matrices to reveal hidden patterns.	Reducing noise and uncovering latent structures.	Requires careful choice of algorithms.
Multivariate Statistics	Analyzing relationships between multiple variables.	Identifying correlated biological processes.	Computationally intensive with large data.
Machine Learning Integration	Using machine learning algorithms for integration.	Integrating omics data with clinical outcomes.	Requires large, labeled datasets for training.

To list various bioinformatics visualization techniques, the types of data they are used for, and their specific use cases (Table 2).

Table 2. Common bioinformatics visualization methods.

Visualization Method	Data Type(s)	Description	Use Case
Heatmaps	Gene Expression	Visualizes expression levels across conditions.	Identifying upregulated or downregulated genes.
PCA	Multi-dimensional data	Reduces dimensionality to visualize clusters.	Visualizing gene expression patterns.
t-SNE	High-dimensional data	Clustering algorithm for non-linear reduction.	Identifying hidden structures in proteomics or genomics data.
Network Graphs	Protein-Protein Interaction	Displays relationships between biological entities.	Analyzing protein interactions.

To summarize the main challenges in bioinformatics data integration and visualization and suggest possible solutions to address these challenges (Table 3).

Table 3. Challenges in data integration and visualization.

Challenge	Description	Impact on Bioinformatics	Solutions/Approaches
Data Heterogeneity	Different data types and formats across sources.	It is difficult to combine and compare datasets.	Standardization and normalization techniques.
Missing or Incomplete Data	Datasets often have missing values or incomplete entries.	Reduces the accuracy and reliability of results.	Imputation methods, data augmentation.
Noise in Data	Variability due to technical errors or biological noise.	Obscures meaningful biological patterns.	Filtering, normalization, outlier detection.
Scalability	Managing large datasets from multiple sources.	High computational cost and time.	Distributed computing, cloud computing solutions.

To present key tools and platforms used for data integration and visualization, summarizing their features and use cases in bioinformatics (Table 4).

Table 4. Tools for data integration and visualization.

Tool Name	Type	Features	Use Cases
Galaxy	Data Integration Platform	Web-based, supports multiple bioinformatics workflows.	Multi-omics data analysis and integration.
Cytoscape	Network Visualization	Visualizes biological networks and molecular interactions.	Gene co-expression and PPI network analysis.
Bioconductor	R-based Package	Comprehensive suite of tools for bioinformatics analysis.	Integrating genomic, transcriptomic, and proteomic data.
Shiny	Interactive Visualization	R-based web applications for real-time data exploration.	Visualizing multi-omics datasets interactively.

CONCLUSIONS

The integration and visualization of biological data are fundamental to advancing our understanding of complex biological systems in bioinformatics. With the ever-increasing volume and complexity of data generated from high-throughput technologies, such as genomics, transcriptomics, proteomics, and metabolomics, researchers are presented with both immense opportunities and significant challenges. Data integration allows for the unification of heterogeneous datasets, facilitating a comprehensive view of biological phenomena, while visualization plays a critical role in making these complex datasets comprehensible, interpretable, and actionable. This study has explored the various techniques used for data integration, including traditional methods, such as concatenation and matrix factorization, as well as more advanced approaches utilizing machine learning and deep learning. These techniques allow researchers to combine data from different omics layers, revealing hidden biological patterns and improving the accuracy of predictions in fields like personalized medicine, drug discovery, and disease research. Visualization tools are indispensable for making sense of high-dimensional, multi-omics data. Techniques, such as heatmaps, PCA, t-SNE, and network-based visualizations help to transform large, complex datasets into intuitive graphical representations. These visualizations facilitate the identification of trends, outliers, and relationships, allowing for more efficient hypothesis generation and hypothesis testing. The development of interactive and dynamic visualization tools has further enhanced the ability to explore data in real time, promoting collaboration and more effective data-driven decision-making.

However, the integration and visualization of biological data still face several challenges. Data heterogeneity, noise, missing values, and the curse of dimensionality all complicate the analysis process. Additionally, scaling these techniques to handle “big data” remains a significant obstacle. Despite these challenges, ongoing progress in computational methods, cloud computing, and machine learning provide promising solutions. Emerging technologies are expected to enhance both the efficiency and accuracy of data integration, while improving visualization capabilities to accommodate larger, more complex datasets.

In conclusion, the successful integration and visualization of bioinformatics data holds the key to unlocking new biological insights, advancing our understanding of health and disease, and accelerating the development of personalized medicine. As bioinformatics continues to evolve, overcoming the existing challenges will pave the way for innovative approaches to data analysis and interpretation, driving further discoveries in life sciences and medicine. The future of bioinformatics lies in the continued refinement of these techniques, with an emphasis on improving computational methods, enhancing visualization tools, and creating more effective, user-friendly platforms for researchers and clinicians alike.

REFERENCES

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nature Genet. 2000;25(1):25–29. doi: 10.1038/75556.
2. Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein-protein interaction networks. J Proteome Res. 2012;11(4):2014-31. doi: 10.1021/pr201211w.
3. Baldi P, Brunak S. Bioinformatics: The machine learning approach. MIT Press; 2001.
4. UniProt C. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–D515. doi: 10.1093/nar/gky1049.
5. Beissbarth T, Speed TP. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 2004;20(9):1464–1465. doi: 10.1093/bioinformatics/bth088.
6. Y. Bengio. Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning. 2009 Jan 1;2(1):1–127. Available from: https://ieeexplore.ieee.org/document/8187120
7. Stein-O'Brien GL, Ainslie MC, Fertig EJ. Forecasting cellular states: From descriptive to predictive biology via single-cell multiomics. Curr Opin Syst Biol. 2021:26:24–32. doi: 10.1016/j.coisb.2021.03.008.
8. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016:17:13. doi: 10.1186/s13059-016-0881-8.
9. Allen GI. Statistical data integration: Challenges and opportunities. Stat Model. 2017;17(4-5):332–337. doi: 10.1177/1471082X17707429.
10. Díez P, Droste C, Dégano RM, González-Muñoz M, Ibarrola N, Pérez-Andrés M, et al. Integration of proteomics and transcriptomics data sets for the analysis of a lymphoma B-Cell line in the context of the chromosome-centric human proteome project. J Proteome Res. 2015;14(9):3530–3540. doi: 10.1021/acs.jproteome.5b00474.
11. Fayyad U, Stolorz P. Data mining and KDD: Promise and challenges. Future Gener Comput Syst. 1997;13(2–3):99–115. doi: 10.1016/S0167-739X(97)00015-0.
12. Mehboob-ur-Rahman TS, Mahmood-ur-Rahman MA, Zafar Y. Bioinformatics: A way forward to explore "plant omics". Bioinformatics-Updated Features and Applications. Croatia: Intech; 2016. p. 203.
13. Ren G, Zhang X, Li Y, Ridout K, Serrano-Serrano ML, Yang Y, et al. Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. Sci Adv. 2021;7(29):eabg2286. doi: 10.1126/sciadv.abg2286.
14. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–754. doi: 10.1016/j.cell.2016.06.017.
15. Aggarwal S, Karmakar A, Krishnakumar S, Paul U, Singh A, Banerjee N, et al. Advances in drug discovery based on genomics, proteomics and bioinformatics in malaria. Curr Top Med Chem. 2023;23(7):551–578. doi: 10.2174/1568026623666230418114455.
16. Patra P, Izawa T, Peña-Castillo L. REPA: Applying pathway analysis to genome-wide transcription factor binding data. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(4):1270–1283. doi: 10.1109/TCBB.2015.2453948.
17. Olsen LR, Campos B, Barnkob MS, Winther O, Brusic V, Andersen MH. Bioinformatics for cancer immunotherapy target discovery. Cancer Immunol Immunother. 2014;63(12):1235–1249. doi: 10.1007/s00262-014-1627-7.

18. Rossini GP, Hartung T. Towards tailored assays for cell-based approaches to toxicity testing. ALTEX. 2012;29(4):359–372. doi: 10.14573/altex.2012.4.359.
19. Fernandez NF, Gundersen GW, Rahman A, Grimes ML, Rikova K, Hornbeck P, et al. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. Sci Data. 2017;4:170151. doi: 10.1038/sdata.2017.151.
20. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90-7. doi: 10.1093/nar/gkw377.
21. Chen S, Qin R, Mahal LK. Sweet systems: Technologies for glycomic analysis and their integration into systems biology. Crit Rev Biochem Mol Biol. 2021;56(3):301–320. doi: 10.1080/10409238.2021.1908953.
22. Graw S, Chappell K, Washam CL, Gies A, Bird J, Robeson MS, et al. Multi-omics data integration considerations and study design for biological systems and disease. Mol Omics. 2021;17(2):170–185. doi: 10.1039/d0mo00041h.
23. Yu XT, Zeng T. Integrative analysis of omics big data. Methods Mol Biol. 2018:1754:109–135. doi: 10.1007/978-1-4939-7717-8_7.
24. McInnes L, Healy J, Saul N, Lukas Großberger. UMAP: Uniform manifold approximation and projection. J Open Source Softw. 2018;3(29):861. doi: 10.21105/joss.00861.
25. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534(7605):55–62. doi: 10.1038/nature18003.
26. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genet. 2003;34(3):267–73. doi: 10.1038/ng1180.
27. Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: An overview. J Biol Res (Thessalon). 2015;22(1):9. doi: 10.1186/s40709-015-0032-5.
28. Platzer A. Visualization of SNPs with t-SNE. PLoS One. 2013;8(2):e56883. doi: 10.1371/journal.pone.0056883.
29. Gutierrez Reyes CD, Alejo-Jacuinde G, Perez Sanchez B, Chavez Reyes J, Onigbinde S, Mogut D, et al. Multi omics applications in biological systems. Curr Issues Mol Biol. 2024;46(6):5777–5793. doi: 10.3390/cimb46060345.
30. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47(D1):D442–D450. doi: 10.1093/nar/gky1106.
31. Cantafio ME, Grillone K, Caracciolo D, Scionti F, Arbitrio M, Barbieri V, et al. From single level analysis to multi-omics integrative approaches: A powerful strategy towards the precision oncology. High Throughput. 2018;7(4):33. doi: 10.3390/ht7040033.
32. Marko NF, Quackenbush J, Weil RJ. Why is there a lack of consensus on molecular subgroups of glioblastoma? Understanding the nature of biological and statistical variability in glioblastoma expression data. PloS One. 2011;6(7):e20826. doi: 10.1371/journal.pone.0020826.
33. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi: 10.1093/nar/gkv007.
34. Diboun I, Wernisch L, Orengo CA, Koltzenburg M. Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC Genomics. 2006;7:252. doi: 10.1186/1471-2164-7-252.
35. Pettini F, Visibelli A, Cicaloni V, Iovinelli D, Spiga O. Multi-omics model applied to cancer genetics. Int J Mol Sci. 2021;22(11):5751. doi: 10.3390/ijms22115751.
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–15550. doi: 10.1073/pnas.0506580102.
37. Zhang Z, Zhao L, Wei X, Guo Q, Zhu X, Wei R, et al. Integrated bioinformatic analysis of microarray data reveals shared gene signature between MDS and AML. Oncol Lett. 2018;16(4):5147–5159. doi: 10.3892/ol.2018.9237.

38. Auwerx C, Sadler MC, Reymond A, Kutalik Z. From pharmacogenetics to pharmaco-omics: Milestones and future directions. HGG Adv. 2022;3(2):100100. doi: 10.1016/j.xhgg.2022.100100.
39. Hill M, Tran N. miRNA interplay: Mechanisms and consequences in cancer. Dis Model Mech. 2021;14(4):dmm047662. doi: 10.1242/dmm.047662.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005 Oct 25;102(43):15545-50.

