Mapping Omics datasets on KEGG Metabolic Pathways

Duarte Velho[1], João Sequeira[2,3], and Andreia Salvador[3]

[1] School of Engineering, Minho University, Campus de Azurém, 4800-019, Guimarães, Portugal
[2] Centre of Biological Engineering, Minho University, Campus de Gualtar, 4710 - 057, Braga, Portugal, direcao@ceb.uminho.pt
https://www.ceb.uminho.pt/

Abstract. This study addresses the challenges of omics data analysis, focusing on the development of improved methods for mapping and interpreting complex networks of biological interactions through omics data. With the increasing complexity and volume of data produced by omics technologies, effective mapping methods are essential to uncover deep biological insights and apply them in clinical and environmental contexts. This project proposes to enhance the KEGGCharter tool by integrating it with interactive functionalities to enable a more dynamic and detailed analysis of gene expression and taxonomy data. Through an original workflow approach, we explore the development of an interactive HTML interface that allows detailed analysis directly on the metabolic maps. This project not only improves the utility of existing omics data mapping tools but also facilitates complex data interpretation and visualization, making a significant contribution to the life sciences and biomedicine by providing a contribution to a richer and more accessible understanding of vast omics data networks.

Keywords: Bioinformatic tools · Metabolic pathways mapping · Metagenomics · Metabolic networks · KEGGCharter · Interactivity.

1 Introduction

1.1 Importance of mapping omics data

Mapping omics data aims to understand the complex network of biological interactions and is playing an increasingly important role in science today.

Microbial communities are made up of bacteria, archaea, fungi, yeasts, eukaryotes and viruses, which often live together in the same habitat. With this size and complexity of microbial communities, the omics data that results from them makes analyzing it a difficult task [23]. Therefore, omics data mapping technologies are of great importance in interpreting omics results.

Laboratory data from the omics sciences has gained widespread interest among researchers in recent years due to its complexity, availability and potential to generate new medical knowledge [24]. Thus, mapping omics data of-

fers attractive opportunities to understand complexity, and gain insight into the underlying biological processes.

Over the past decade, omics data mapping has emerged as a valuable aid to understanding the data generated by various "omics" technologies. As a result, several robust software tools as well as web interfaces for meta-omics analysis have been developed to support pathway analysis for genomic and proteomic studies [26].

1.2 Current challenges in mapping omics data

Software developed for omics and meta-omics analysis, together with knowledge bases that include information on genes, proteins, taxonomic and functional annotation, among other types of information [23], have become powerful resources for omics analysis. In this way, the new panorama we are facing requires the development of new software tools [3], which automatically convert raw data into complete information through metabolic maps.

Although several bioinformatics resources are available for meta-omics analysis, many of them require significant computational knowledge [23] and a large computing capacity. Web interfaces are easier to use, but often have difficulty dealing with large data files [23].

One of the great advantages of some omics data mapping software tools is the interactive identification feature, which allows users to interact with specific elements of a pathway for further exploration or to obtain additional data on genes, specific proteins or even taxonomy in order to identify evolutionary relationships between species. This feature is essential for researchers who want to understand their experimental data in more detail.

1.3 Review of omics data analysis tools used in mapping pathways

With the availability of metabolic network pathway diagrams and an ever-increasing volume of omics data to visualize, various tools have been developed to facilitate the interpretation of functional annotation results and to represent the genes or proteins identified in metabolic pathways.

KEGGCharter [23] is a command-line implementation of the KEGG Pathway mapping service, also obtaining additional KOs and EC numbers through the methods available in BioPython to access the KEGG API. KEGGCharter takes as input a table (TSV or EXCEL), which contains KEGG IDs, KOs or EC numbers, represents KOs identified in metabolic maps and includes information on differential gene expression. When data from more than one organism is uploaded, KEGGCharter links the function to the taxonomic identification, which can be visualized on the maps. The differential expression of genes/proteins can be visualized on metabolic maps, showing mini heat maps.

KEGG Mapper [12] is a collection of pathway mapping tools, BRITE and MODULES. It provides a comprehensive database that integrates systemic functional information with genomic and chemical data. KEGG Mapper [12] offers interactive functionalities, enabling users to explore detailed information on

molecular interactions and functions. Its tools, such as "Search&Color Pathway", make it possible to map gene expression data, indicating positive or negative regulation.

MetPA [26], is a tool designed for the visualization of metabolomic data. It uses packages such as KEGGgraph, Graphviz and ImageMagick to render detailed metabolic networks, available in a web interface that supports zooming without loss of quality and dynamic manipulation. The tool covers 11 models of organisms. The information is visualized as a network in a web interface, providing a robust and easy-to-use metabolic pathway analysis environment [26].

DAVID [4] is a web-accessible software that integrates genomic functional annotations with intuitive graphical summaries, simplifying the annotation and rapid summarization of data according to shared categories, such as Gene Ontology and protein domains.

CellDesigner 3.5 [7], is a modeling tool for biochemical and gene-regulatory networks, using standardized graphical notation and SBML (Systems Biology Markup Language). It has an intuitive interface and supports various biological identifiers. It can include differential gene expressions in its models and exports models in graphic formats such as PNG and SVG [7].

Similarly, GenMAPP [22] is a tool that allows users to visualize and analyze genome-scale data on biological pathways. Implemented in Visual Basic 6.0, it supports multiple annotations of genes and species and allows the customized creation of databases for a potentially unlimited number of species. This tool integrates ETL processes for extracting data from public resources such as Ensembl, Entrez Gene and Affymetrix, supporting a wide range of identifiers such as Ensembl gene IDs, UniProt IDs, Entrez Gene IDs and Affymetrix probe set IDs. A unique feature is the dynamic coloring of genes in lanes based on user-defined criteria. GenMAPP [22] allows the visualization of differential gene expression data.

KGML-ED [17], allows dynamic visualization, interactive navigation and editing of KEGG road diagrams. It allows the editing and creation of new routes, facilitating the integration of user-specific data. KGML-ED [17] offers semi-static and dynamic visualization techniques for enhanced road analysis. It supports KGML identifiers for genes, allowing detailed mapping and annotation within the KEGG framework.

MetaCore™ [5] is a web platform that analyzes high-throughput molecular data and integrates it for visualization in the context of metabolic maps. With paid access and complexity for new users, it uses Perl, HTML/JavaScript and Flash Player Plug-in. It accepts identifiers such as LocusLink, SwissProt, RefSeq and Unigene. Interactivity is possible. It supports differential gene expression analysis and accepts various data input and output formats, including XML, JSON, CSV, TSV, PDF, PNG, JPEG, HTML, XLS and XLSX.

Pathway Tools [14] is designed for simulation and visualization of integrated collections of genomic data. It allows metabolic reconstructions and predictions, as well as visualization of regulatory interactions and analysis of omics data [14]. Over the course of its versions, Pathway Tools has improved its capabilities and

expanded its database from 800 genome/pathway databases (PGDBs) [13] to more than 20000 [15]. Learning the tool can be steep for new users. Pathway Tools [14] maintains interactivity and compatibility with different data input and output formats, such as GenBank, SBML and BioPAX.

The iPath tool [27] is a web tool for visualizing and analyzing cellular pathways that provides interactive maps for central metabolism. It offers an interactive and user-friendly interface for exploring a wide range of metabolic pathway data, with the ability to map a variety of identifiers, such as KEGG, and customize maps with data such as differential expression. Over the course of updates, iPath has expanded its dataset and functionalities, including new modules for KEGG pathways, reactions and species [1]. The latest version supports interactive analysis [1].

PathVisio [19] is a software tool for visualizing, editing and analyzing biological pathways. Although it requires manual data entry, which can be time-consuming, PathVisio is enhanced by its ability to interact with other scientific tools, such as Cytoscape and Eu.Gene.

Pathview [18] is a suite of tools for integrating and visualizing metabolic pathway-based data, mapping user data onto relevant pathway graphs for easy analysis and interpretation. As part of the Bioconductor project, it integrates with several R packages for comprehensive data analysis and supports a wide range of identifiers for genes/proteins and compounds/metabolites [18]. It is interactive, dealing with datasets of different scales and complexity, and compatible with more than 2000 species. It supports the visualization of differential gene expression.

Pathway Tools [21] is a Java-based application that provides a visual representation of an organism's biochemical network, automatically generated from a Pathway/Genome Database (PGDB). The tool is designed to work with whole organism datasets, integrating with Omics Viewer to overlay gene expression data and other quantitative data. Enhancements over the years include interactive web and desktop views, semantic zoom support, poster generation and colorful, animated views for time series data [20].

Reactome Knowledgebase [6], is an online database that details cellular processes in an ordered network of molecular transformations. It allows interactive visualization of omics data on pathway diagrams and offers export of pathway diagrams for further analysis and customization by the user [8].

Looking at the overview of the tools described, KEGG Mapper [11] and Reactome [8] offer in-depth functionalities for exploring genomic and proteomic data, allowing detailed mapping of pathways with visual adjustments based on gene expression, differing from MetaCore™ [5] which, despite providing detailed analysis, requires a paid subscription and has a steeper learning curve due to its complexity. Tools such as Pathway Tools [21] and MetPA [26], while also focused on metabolic pathways, differ in their approaches; Pathway Tools [21] allows for detailed simulations and predictions within an integrated environment, while MetPA [26] specializes in visualizing metabolomic data with a highly interactive interface. Tools such as Pathview [18] and PathVisio [19] facilitate

the integration and visual analysis of pathway data, with PathVisio [19] being enhanced by its ability to interact with other scientific tools such as Cytoscape. In contrast, DAVID [4] and KEGGCharter [23] are more focused on functional annotation and data summarization, with DAVID [4] simplifying the visualization of categories such as Gene Ontology and KEGGCharter [23] allowing direct mapping via the command line, which makes this tool interesting from the point of view of user customization. This diversity of tools offers users a wide range of options depending on their specific analysis and visualization needs.

The CellDesigner [7] and KGML-ED [17] tools are similar but do not represent the taxonomy, which makes them invalid for anyone who wants to relate metabolomic data to the organism's taxonomy.

With the exception of KEGGCharter [23], most of the tools described have interactivity, which is a great advantage when it comes to metabolic pathway mapping tools.

One way of introducing interactivity into the tool would be to implement JavaScript to create interactive elements. A practical example of this would be KronaTools (Click here to visit the git hub), which generates Krona charts that offer a powerful and intuitive way to visualize and explore hierarchical data, with interactivity that allows users to analyze the data at different levels of detail and customize the visualization as needed. Of the tools described, only Reactome [8], MetaCore™ [5] and iPath [27] use JavaScript tools to provide interactivity to the mapping tools.

(Click here to visit the Table)

Fig. 1. The available analytical tools and web servers designed for mapping pathways data, including their functions, advantages, disadvantages, supported languages, types of identifiers, interactivity, sample representation, taxonomy considerations, differential expression analysis capabilities, and input/output formats.

2 Context and importance/relevance of the project

At the forefront of bioinformatics, the project under analysis is part of the emerging field of metagenomic approaches, an area recognized for its potential in deciphering the complexity and dynamics of microbial ecosystems, as well as its relevance in the study of disease evolution. In this context, the ability to map omics datasets onto metabolic pathways represents an indispensable tool in interpreting complex interactions in genetic and metabolic networks, as the analysis of omics data is quite complex and is greatly simplified with the introduction of interactive metabolic maps that provide the user with a personalized experience. Metagenomics opens the door to a more comprehensive understanding of microbial ecology, enabling significant advances in science and medicine, particularly in the identification of new biomarkers and therapeutic targets.

With the contribution proposed in this project, to improve KEGGCharter by introducing interactivity into the graphs generated by the tool, users are being given the opportunity to integrate the many advantages of KEGGCharter into an interactive element and make it easier to draw conclusions from their omics data on the metabolic map.

3 Motivation

3.1 KEGGCharter

KEGGCharter [23] is a tool that represents the omics results, including differential gene expression, in the KEGG metabolic pathways. In addition, it shows the taxonomic assignment of the enzymes represented, which is particularly useful in metagenomic studies in which several microorganisms are present. Currently, KEGGCharter shows the differential expression of genes and the taxonomic assignment of genes as separate entities on the maps. This separation limits the usefulness of directly associating microbial identities with gene expression data in a unified visual context. The graphical outputs that this tool produces are static, providing the user only with the information contained therein. One way to counteract this disadvantage would be to integrate interactivity into the graphics generated by KEGGCharter, in order to give the reader the possibility of having a personalized analysis of the information, with the integration of taxonomy data, differential expression, among others, simultaneously.

3.2 Objective of the study

The aim of the project is to modify the KEGGCharter to provide interactive graphical outputs, giving the user the possibility of viewing the information at different levels, according to their personalization. The KEGGCharter graphics will be expanded to include multi-level representation of gene expression information. The work will involve the development of new interactive representations on top of the graphs currently produced, to allow more information to be included in the metabolic maps.

4 Methodology

To tackle the problem of the lack of interactivity in the graphics generated by KEGGCharter output, several JavaScript functionalities will be added to the metabolic maps. For each function in each metabolic map, a detailed HTML page will be created, containing specific information about the selected gene or enzyme, including gene expression data, taxonomic assignment and other relevant information. Within this page there will also be a link that takes the user to the corresponding page on KEGG, based on the K numbers contained in each box of the graph generated by KEGGCharter.

There are a number of Javascript packages for generating interactive elements that may be employed in this work. The graphical outputs generated by KEGGCharter are generated in png format. To be able to access the elements of the graph and thus generate HTML pages, the png format have to be converted to svg. The Potrace package may be used for this purpose. Three.js is a lightweight, multi-browser JavaScript library/API used to create and display animated 3D graphics in a web browser, and may be used in this work, in conjunction with the HTML5 canvas element, SVG or WebGL. Another interesting package is D3.js, which is a JavaScript library for cre-ating interactive visualizations of document data using HTML and SVG.

Acknowledgments. I would like to express my sincere gratitude to the faculty and staff of the School of Engineering and the Centre of Biological Engineering at Minho University for their invaluable support and guidance throughout the development of this project. Special thanks to my familie and friends for their patience and encouragement. This work was made possible by the resources provided by the Centre of Biological Engineering, for which I'm profoundly grateful.

References

1. Darzi, Y., Letunic, I., Bork, P., and Yamada, T.: iPath3.0: interactive pathways explorer v3. Nucleic Acids Research **W510-W513** (2018). https://doi.org/10.1093/nar/gky299.
2. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., and Krawetz, S.A.: Global functional profiling of gene expression. Genomics **81**, 98–104 (2003). https://doi.org/10.1016/S0888-7543(02)00021-6.
3. De Filippo, C., Ramazzotti, M., Fontana, P., Cavalieri, D.: Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Briefings in Bioinformatics **13**(6), 696–710 (2012). https://doi.org/10.1093/bib/bbs070.
4. Dennis Jr, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., & Lempicki, R.A.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology **4**(9), R60 (2003).
5. Ekins S., Nikolsky Y., Bugrim A., Kirillov E., Nikolskaya T. Pathway mapping tools for analysis of high content data. In: Taylor D. L., Haskins J. R., Giuliano K. A. (eds.) High Content Screening: A Powerful Approach to Systems Cell Biology and Drug Discovery, vol. 356, pp. 319–350. Humana Press, Totowa, NJ (2007).
6. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Duenas Roca, C., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P.: The Reactome Pathway Knowledgebase. Nucleic Acids Research **46**(D1), D649–D655 (2018). https://doi.org/10.1093/nar/gkx1132.
7. Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., Kitano, H.: CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. Proceedings of the IEEE **96**(8), 1254–1265 (2008). https://doi.org/10.1109/JPROC.2008.925458.
8. Haw, R., Hermjakob, H., D'Eustachio, P., & Stein, L. (2011). Reactome pathway analysis to enrich biological discovery in proteomics data sets. Proteomics, **11**(18), 3598–3613.

9. Jin, L., Zuo, X.-Y., Su, W.-Y., Zhao, X.-L., Yuan, M.-Q., Han, L.-Z., Zhao, X., Chen, Y.-D., Rao, S.-Q.: Pathway-based Analysis Tools for Complex Diseases: A Review. Genomics Proteomics Bioinformatics **12**(5), 210–220 (2014).
10. Kanehisa, M., Sato, Y.: KEGG Mapper for inferring cellular functions from protein sequences. Protein Science **29**, 28–35 (2020). https://doi.org/10.1002/pro.3711.
11. Kanehisa, M., Sato, Y., Kawashima, M.: KEGG mapping tools for uncovering hidden features in biological data. Protein Science **31**(1), 47–53 (2022). https://doi.org/10.1002/pro.4172.
12. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research **40**(D1), D109–D114 (2012). https://doi.org/10.1093/nar/gkr988.
13. Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., Caspi, R.: Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in Bioinformatics **11**(1), 40–79 (2009). https://doi.org/10.1093/bib/bbp043.
14. Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I.M., Caspi, R.: Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Briefings in Bioinformatics **17**(5), 877-890 (2016). https://doi.org/10.1093/bib/bbv079.
15. Karp, P.D., Midford, P.E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W.K., Subhraveti, P., Caspi, R., Fulcher, C., Keseler, I.M., & Paley, S.M.: Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Briefings in Bioinformatics **22**(1), 109–126 (2021). https://doi.org/10.1093/bib/bbz104.
16. Khatri, P., Draghici, S., Ostermeier, G.C., Krawetz, S.A.: Profiling Gene Expression Using Onto-Express. Genomics **79**(2), 266–270 (2002). https://doi.org/10.1006/geno.2002.6698.
17. Klukas, C., Schreiber, F.: Dynamic exploration and editing of KEGG pathway diagrams. Bioinformatics **23**(3), 344–350 (2007). https://doi.org/10.1093/bioinformatics/btl611.
18. Luo, W., Brouwer, C.: Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics **29**(14), 1830–1831 (2013). https://doi.org/10.1093/bioinformatics/btt285.
19. van Iersel, M.P., Kelder, T., Pico, A.R., Hanspers, K., Coort, S., Conklin, B.R., & Evelo, C.: Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics **9**, 399 (2008). https://doi.org/10.1186/1471-2105-9-399.
20. Paley, S.M., Karp, P.D.: The Pathway Tools cellular overview diagram and Omics Viewer. Nucleic Acids Research **34**(13), 3771–3778 (2006). https://doi.org/10.1093/nar/gkl334.
21. Rahman, S.A., Advani, P., Schunk, R., Schrader, R., Schomburg, D.: Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics **21**(7), 1189–1193 (2005). https://doi.org/10.1093/bioinformatics/bti116.
22. Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J., Conklin, B.R., Pico, A.R.: GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics **8**, 217 (2007).
23. Sequeira, J.C., Rocha, M., Alves, M.M., Salvador, A.F.: UPIMAPI reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets. Comput. Struct. Biotechnol. J. **20**, 1798–1810 (2022).

24. Toussaint, P. A., Leiser, F., Thiebes, S., Schlesner, M., Brors, B., Sunyaev, A.: Explainable artificial intelligence for omics data: a systematic mapping study. Briefings in Bioinformatics **25**(1), 1–16 (2024). https://doi.org/10.1093/bib/bbad453.
25. Werner, T.: Bioinformatics applications for pathway analysis of microarray data. In: Current Opinion in Biotechnology **19**, pp. 50–54 (2008). https://doi.org/10.1016/j.copbio.2007.11.005.
26. Xia, J., & Wishart, D.S.: MetPA: a web-based metabolomics tool for pathway analysis and visualization. In: Bioinformatics Applications Note, Vol. 26, No. 18, pp. 2342–2344. Oxford University Press (2010). https://doi.org/10.1093/bioinformatics/btq418.
27. Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., Bork, P.: iPath2.0: interactive pathway explorer. Nucleic Acids Research **39**(suppl_2), W412–W415 (2011). https://doi.org/10.1093/nar/gkr313.