A Computational Biology Framework

Creating a platform for biomedical engineers to efficiently do their research

T.P.A. Beishuizen (0791613)
Biomedical Engineering - Computational Biology
Data Engineering - Information Systems
Eindhoven, University of Technology
Email: `t.p.a.beishuizen@student.tue.nl`

March 5, 2018

Contents

1 Introduction

At the Computational Biology department (cBio) of Biomedical Engineering (BME), many requests were made to analyse gathered data. This data usually stemmed from research in hospitals, but could also be from other BME groups and publicly available. Currently a standard was missing to efficiently analyse those data sets. With the vast number of data sets that were available, such a standard in the form of a framework on data analysis would be valuable. It would speed up projects and give them a higher chance to succeed the goal, due to improved efficiency. Before a framework could be made however a research must be done on all aspects that influence data analysis.

First an extensive background on important topics for such a framework would be discussed. Four different parts were explained why they are important for the creation of such a framework. These parts are: biomedical data (data used for analysis), data analysis goal (how does a goal influence the choice of analysis), data analysis tools (which ones are usable) and at last biomedical knowledge (what engineers of BME and third parties already know about data analysis). After the extensive background research, a research question would be formulated with several sub-questions for each of the four parts and a hypothesis as an answer for each of the four questions.

2 Biomedical Research Projects

Biomedical engineering could be seen as a specific part of engineering with a wide variety of topics. These topics were theoretical, non-experimental undertakings, but also state-of-the-art applications. Not only research and development was used, but also implementation and operation. Combining all of these different parts in one definition was hard.[1] For this project, the focus was mainly on research and development, also known as knowledge discovery.[2]

When a biomedical engineer started a project, at the start usually only a data set and the research goal were known. To achieve that goal from the data set, several different aspects influenced the project's course and development. At first obviously the data itself was a big part of such an influencer as the research was restricted to limitations from it. Examples of such restrictions were multidimensionality, set size, data heterogeneity, missing feature values and population handling. The other obvious influencer was the main research goal. Since the biomedical engineer wanted to achieve a certain goal, the approach outcome had to match that goal for the research to be successful. Most goals were focused around either data mining, extracting relations from available data, or modelling, creating a model within data features. A third influencer is the availability of data analysis tools. The steps to take from data to goal did not only include an approach, but also a tool to execute it. The choice of a certain tool had a big impact on the project, as each one of them had its own advantages and disadvantages. The two most well known tools within BME were MATLAB and Python, however some engineers had used R, Java or C++ and there were still other possibilities. A last big influencer was the biomedical knowledge. What experience the scientist already had with similar projects could greatly influence the choice of approach and framework. Knowledge of the supervisor and publicly known information on the research subject from books and articles also influenced the approach, as already known outcomes did not have to be researched again.

Previous research projects on data mining had called for a model how to retrieve patterns from data collections. Frameworks to do that had been proposed effectively, usually with a number of steps.[3] These suggested frameworks did not specifically fit the cBio group though for being too broad[4] or being too specific.[5] A customized framework would be very beneficiary.

2.1 Biomedical Data

A big aspect of choosing how to set up the data analysis was the data itself. The amount of data in the biomedical world was growing at an enormous rate, faster than biomedical engineers could analyse. Due to this rapid growth being uncontrollable, several additional challenges arose, aside being more than the biomedical world could handle. These challenges were mainly focused around data volume, dimensionality, complexity, heterogeneity and quality.[6, 7]

Collecting data because it was possible could make data sets bigger than needed. Both in number of instances and features, data sets could be harder to understand or analyse when more is available.[6] This volume problem usually was tackled by taking sub-populations of the complete set. Sub-sets could either be focused around a part of the population (gender, age, race) or taken at random to still represent all of it. Due to the efficiency of analysis techniques and the rise in computational speed of servers[8], volume on its own became less of an issue. Volume does however become an issue when combining with heterogeneity and quality. [9, 10]

Not all data sets had a high number of instances that cause a big data volume. Sometimes there were relatively few instances, while the number of features was proportionally high.[11] Usually many of those features were not relevant enough for the research, however were still used for testing. Trying to remove features that were not important would greatly help finding relations between the others and create more knowledge about the research topic. Lowering the number of features also made the data volume go down, so analysis should be easier. Mainly an optimal features set should be selected to obtain the best results.[12]

Another way to approach a high number of features for a low number of samples would be that the number of samples were too low. If data was gathered for only a couple patients, results would hardly ever be consistent. Most analyses required more samples and would give biased results because of that. There were two ways of dealing with this low number of samples, the first one being generating more samples**??** and the second one using the samples highly efficient[13, 14]. Most likely a combination fo both would give the best results.

Biomedical data could also be very complex. Useful results could be present, however it was very hard to obtain it. Examples of complex data were images, several biomedical signals and temporal data. Details of the useful results that were present in images was for example very hard to detect, the temporal data could vary quite much over time and the biomedical signals could be hard to combine with static biomarkers.[15] This aspect could benefit from exchanging knowledge with other research areas that specialize in mining of those complex data sets.[9, 16]

The biggest challenge encompassed aligning different data sets. No standard for data sets was available and therefore data sets differed greatly from each other. Data was weakly structured or even unstructured[10] and variables were processed differently due to other protocols or the collectors' preference of representation.[17] Also the variety of data was hard to combine when sources were fundamentally different. When parts of the data were images, another part was a table from the laboratory and a third part was textual remarks of the doctor, standardizing merging those three was much harder than merging three lab sets. Those merges were also very prone to errors, as imprecisions could be vastly different between those data sets. No tool worked directly with these raw data sets and preprocessing almost definitely had to occur beforehand.[3, 9]

A last challenge was about data quality. The data was usually gathered by doctors and laboratory workers. Since the data was manually gathered by humans, the data had a relatively high error rate. The data could be quite noisy, values could be inconsistent, wrongly entered or even missing.[3] Not only human errors caused the data quality to drop, but the heterogeneity, as well. Two hospitals could have different protocols for the same treatment and sample different biomarkers for that protocol. Due to that difference, biomarkers could be missing for some of the entries. The time of data gathering was also a big factor as some biomarkers changed greatly over time. The databases were usually also built for financial purposes and not for research, which could hurt the quality.[15]

These challenges within the data were greatly discussed.[16] Many proposals to tackle them were made, however none was actually widely adopted, yet, as a global standard for databases. Also, with the uncontrolled growth in biomedical data, it would become hard to have such a standard recognised all over the world.[17, 18, 19, 20, 21, 22]

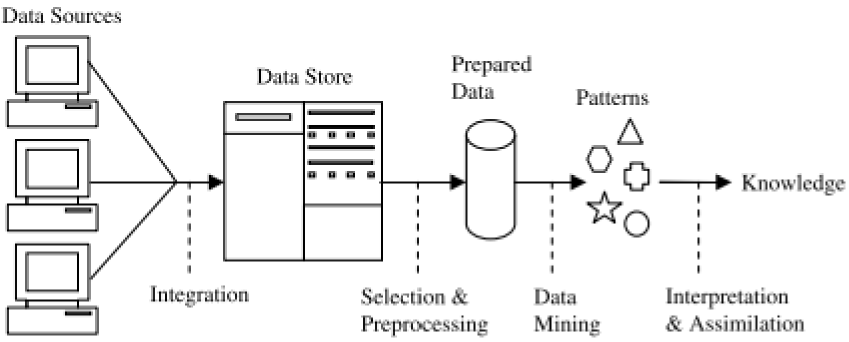

Figure 1: A schematic overview how a project involving data mining was done. Multiple data sets were integrated in one database. Feature selection and preprocessing took place to prepare the data. Then the data was mined to find patterns. These patterns were then interpreted and assimilated to discover knowledge on the subject.[2]

2.2 Research goal: Data Mining

A second important part was the research goal. Available data was analysed with a certain goal in mind. This goal could differ greatly for multiple projects and of course heavily influenced the data analysis approach. Two major aspects were present for data analysis, data mining and modelling. For this project, the focus was on data mining.

Data Mining gained multiple definitions over the years. The most adopted definition was the following: "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."[23] As could be read in this definition, the main goal was find new and useful insights and patterns through datasets, that could be used in further decisions or hypotheses.[6, 24] It was one of the links that brought data and knowledge together.[15] A data mining project followed a certain layout. (Figure 2.2), which in the end lead to further knowledge that aided the medical world.[2]

Since biomedical data was a wide scope, data mining had several specialisations in different directions. An example would be text data mining, used to find related articles with websites such as Pubmed and Google Scholar. These articles could mention similar genes, diseases or proteins and give additional information about them. Another example would be microarray data mining. This type of data mining focused on extracting entities and pathways that defined a disease or other phenotype. Two other data mining types were proposed. One focused on extracting useful information out of mass spectrometry data points, called proteomic data mining. A second focused more on a molecular level and how these molecules would affect different cell types.[25]

Whichever splits were made in the data mining term for different areas in biomedical data analysis, the techniques were mostly based on statistical analysis and machine learning. A discussion of these two types had been made earlier report[1].

2.3 Data Mining Tools

When applying several data mining algorithms to extract results from a data set, a certain tool had to be used that facilitated the use of those algorithms. Several basic tools were available and widely used[1]. Since biomedical data had specific characteristics, separate tools or frameworks consisting of those tools were made. These were usually made to create a database with a certain lay-out to put it in or to collect multiple algorithms and alter them specifically for biomedical cases. Many times both were connected as well, for optimized use.

[1] Biomedical Data Analysis - *Tim Beishuizen*

To store biomedical data several projects were started. Famous project examples were the Human Genome Project (HGP) that tried to sequence the entire human genome[26] or the Human Connectome Project (HCP) that focused on the brain connectivity.[27] Both consisted of an enormous amount of data that needed to be stored and a tool was created for use in these project, for instance the ConnectomeDB for the HCP.[28] Similar projects resulted in biomedical databases, such as KEGG[29], Reactome[30], NCBI[31], GeneCards[32] and HMDB[33]. These databases stored data about genes, metabolites and pathways and were used by numerous scientists to completely understand those. More databases were made or are in development, all of them for a specific domain in biomedical research.

Several examples of tools and frameworks of biomedical data mining were available. A framework designed specifically for biomedical data was ImageJ[34]. ImageJ could be used for biomedical image analysis and consisted of image processing and analysis techniques. Another example would be Genesis[35], a platform for gene expression analysis based on Java and consisting of various preprocessing and clustering algorithms. Similarly the tools GeneCodis3[36], FatiGO[37] and [38] were created to interpret results from genomics generated data. Similar tools were made for metabolomics, such as Metscape 2[39] and MetaboAnalyst[40]. The tool G*power3[41] focused more on statistical analysis of biomedical data.

One immediate issue with all of theses existing databases would be that hardly any framework or tool could be found for synchronisation of these databases. Their use was limited to a specific area only without any possibilities to switch in between. An all encompassing framework was not made. Therefore multiple scientists made their work specifically for their specific area. New scientists had a hard time using the newly created tool and tried to create a new one themselves. They would benefit of a new framework that combined aspects of the available tools.[42]

2.4 Biomedical Knowledge

A last important aspect for data analysis was the available knowledge on the topic. Numerous scientists had done research on various biomedical topics. The knowledge gathered from these topics were available in several different ways.[16] Usually this information could be found in a textual manner, in books, papers and articles describing the project results. Other information could be found in the earlier mentioned databases, tools and designed algorithms. A third type of knowledge was implicit, skills and information scientists are expected to have or should be trained for if not known with it.[43] All of these were interesting in their own ways.

Literature knowledge should be the basis for all research. The knowledge found on several different research areas could result in new insights and ideas for unexplored parts. The projects generated this way became knowledge-driven and had hypotheses based on gathered information.[16] Not only at the start of research projects knowledge was used, as it should be a guide throughout the whole process. In the final project steps their results should be used to improve the existing knowledge, so it could be used in new projects.[43]

When looking at projects on implementation level, more interesting data knowledge could be gathered from databases, tools and algorithms. These three could be used to test and answer hypotheses derived from textual knowledge. Mainly these were used for data analysis as known data was gathered from databases, tools were used for their analysis properties and algorithms were implemented, all to improve the project results. When considering data mining and its tools (subsection 2.2 and 2.3), everything known plays a big role, too.[16]

A third knowledge type was the implicit knowledge of the scientists conducting a research project. These scientist could have a big difference in background and therefore focus on different goals, tools or algorithms. Scientists should be familiar with using data mining tools to be able to use them, both in a general way as well as for specific tools. Frameworks were made with a certain type of scientist in mind as well, making it important that the desired type of scientist was able to use it, too.[43]

These types of knowledge could cause several issues. Projects would not be conducted properly if literature did not back up initial points of view and assumptions. On the other hand data knowledge that stemmed from other research projects could be used for continued improvement in the projects, that could not be done if this type of knowledge was not around. At problems could arise when scientists had different implicit knowledge. A physician at the hospital would not have the same knowledge as a professor in data mining, but would have a better understanding of the usefulness of the results.

3 Research Question

The main goal was to create a framework for more efficient data analysis. Based around the four main influencers of data analysis, a main research question was made and divided in five separate sub-questions. The main question would be:

What aspects are of importance to include in a framework for Biomedical Engineers to do data analysis more efficient?

As discussed earlier the main question could be divided in five different sub-questions. These five sub-questions would have their separate hypothesis. The five questions were:

Data driven How does the available data influence the choice for a certain data analysis approach?

Target driven How does the research goal influence the choice for a certain data analysis approach?

Tool availability driven Which available research tools should be included in this framework?

Tool extension driven In which instances is a tool extension required for completion of the framework?

Knowledge driven How must both availability- and lack of knowledge be included in the framework?

3.1 Data Driven Hypothesis

A major part in a data driven analysis approach was the presence of challenges with preprocessing the data. Data volume, dimensionality, complexity, heterogeneity and quality challenges all were issues to be tackled, some of them more interesting than others. The most approachable three challenges would be the data volume, dimensionality and quality challenges. These three problems were well known defined problems with various ways to deal with them, however no perfect answer existed yet. Data dimensionality could be tackled by various feature selection and reduction algorithms, such as principal component analysis, significance testing and clustering. Low data volume could be tackled by various data generation algorithms based on nearest neighbours and interpolation. At last data quality can be improved by handling for example missing values with interpolation and outliers with outlier detection.

3.2 Target Driven Hypothesis

To reach a certain goal, several steps have to be taken. While the steps could be vastly different, it rarely would not start with a global analysis to localize possible issues and interesting features in the data. A statistical analysis should be very beneficial for doing such a global analysis and a possibility for adding that to the framework would be a great benefit.

Other steps were less clear. Follow-up steps after a global data analysis would be to try out several basic techniques known to achieve that goal. After computing the results, they could be used to revise the technique for better results or disregard it completely. This more specific data mining step would include trying out several machine learning techniques or better feature selection.

The final steps to achieve the goal would be very personalized for every project. To approach these final steps numerous things should be considered and done already in previous steps. This would mean that the scientist itself would have the best idea how to achieve that goal with the knowledge gathered from the first couple of steps. Therefore it would not be useful to guide someone in this part of the data analysis.

3.3 Tool Availability Driven Hypothesis

Tools to be included in the final framework should be focused around the data driven and target driven desired techniques. This would mean that it should include tools for feature reduction and selection, data generation, missing value and outlier handling, global analysis and machine learning. On top of that the tools should be able to communicate easily, as they might be used in sequence.

The package scikit-learn from python provided already several algorithms for this problem and would be a good start for the framework. Several feature selection and machine learning algorithms were present and easily useable. A very welcome addition to scikit-learn was Tree-based Pipeline Optimization Tool (TPOT) that made finding the best machine learning algorithm with the best hyperparameters automated. TPOT used both algorithms from scikit-learn and algorithms made by the user. These two would be good tools for a successful data mining project.

Both scikit-learn and TPOT were packages for Python. Another Python package that would be helpful for the project would be the SciPy package. SciPy was very useful for statistical analysis, due to the broad level of available statistical algorithms in it. Those algorithms would be good for global analysis and outlier detection as well as possible feature reduction. Anaconda was a Python distribution that could help you automatically download multiple packages, among others scikit-learn, SciPy and the widely used NumPy. This Anaconda would therefore be a welcome addition, too.

3.4 Tool Extension Driven Hypothesis

With the mention of Scikit-learn, TPOT and SciPy, most aspects desired in the framework were at least partly covered. These three most likely were not perfect for every data set though and might need more programming to better suit the framework. Also several data driven preprocessing algorithms to TPOT would be helpful, making TPOT more suitable for biomedical data.

With the mentioned tools a good base for most of the desired parts in the framework were mentioned. The only two aspects that were not discussed, yet, would be the volume generation and missing value handling. Both of these and possible more algorithms for the other aspects could be programmed to show more variety in preprocessing and data mining.

Some statistical methods were available for analysis in SciPy. These statistical methods still had to be called for privately for SciPy and no selection was made for an initial global analysis. Therefore a new method for doing that global analysis should be made, that borrowed methods from SciPy to give the use insight in the dataset. This would mean it needed basic statistical methods, such as descriptive statistics computation, significance testing and multicollinearity testing.

At last numerous online databases were available with both genomic and metabolic data in them, such as NCBI and KEGG. Some people made some possible integration packages for these databases, however these were not widely known or used. A possible tool extension for the framework would be a possibility to download specific parts of these databases that would contain useful information, so these can be used for further investigation of the genomic or metabolic data.

3.5 Knowledge Driven Hypothesis

The framework would be mainly created for the cBio group. Since this group mainly consisted of either experienced employees that were already doing a research projects or less experienced students, the best focus for this framework would be biomedical students. They would benefit the most for such a framework that could help them at the start of their projects. The scope of the framework would then be people that had some minor programming skills and would need guidance for better understanding and efficiency.

Since inexperienced programmers would not know immediately how to program in possibly a different language, a graphical user interface (GUI) could be of help as well. This GUI could help mainly in the global analysis, but maybe also in running some initial calculations with TPOT.

When using this GUI, programming would be less important as a skill and therefore people would not need to figure out how the framework was designed. For programmers that needed to use more specific methods, it would still be possible to use it as packages for their own project.

To understand everything without too much difficulty a user manual should be made. This user manual should include (publicly available) data sets and show how the framework should be used exemplary with these data sets. This manual and explanations would make the framework more accessible for both inexperienced and experienced programmers and would help them understand it quicker.

References

[1] J. D. Bronzino and D. R. Peterson, *Biomedical engineering fundamentals.* CRC press, 2014.

[2] M. Bramer, *Principles of data mining*, vol. 180. Springer, 2007.

[3] K. J. Cios and G. W. Moore, “Uniqueness of medical data mining,” *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 1 – 24, 2002. Medical Data Mining and Knowledge Discovery.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *et al.*, “Knowledge discovery and data mining: Towards a unifying framework.,” in *KDD*, vol. 96, pp. 82–88, 1996.

[5] K. J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma, “A knowledge discovery approach to diagnosing myocardial perfusion,” *IEEE Engineering in Medicine and Biology Magazine*, vol. 19, no. 4, pp. 17–25, 2000.

[6] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh, *Medical informatics: knowledge management and data mining in biomedicine*, vol. 8. Springer Science & Business Media, 2006.

[7] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in Bioinformatics*, p. bbx044, 2017.

[8] D. Blythe, “Rise of the graphics processor,” *Proceedings of the IEEE*, vol. 96, no. 5, pp. 761–778, 2008.

[9] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser, *On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics*, pp. 117–140. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[10] A. Holzinger and I. Jurisica, *Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions*, pp. 1–18. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[11] W. Dubitzky, M. Granzow, and D. P. Berrar, *Fundamentals of data mining in genomics and proteomics.* Springer Science & Business Media, 2007.

[12] Y. Peng, Z. Wu, and J. Jiang, “A novel feature selection approach for biomedical data classification,” *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 15 – 23, 2010.

[13] L. J. Van’t Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen, *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer,” *nature*, vol. 415, no. 6871, p. 530, 2002.

[14] D. A. Roff and P. Bentzen, “The statistical analysis of mitochondrial dna polymorphisms: chi 2 and the problem of small samples.,” *Molecular biology and evolution*, vol. 6, no. 5, pp. 539–545, 1989.

[15] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, “Data mining in healthcare and biomedicine: A survey of the literature,” *Journal of Medical Systems*, vol. 36, pp. 2431–2448, Aug 2012.

[16] R. Bellazzi, M. Diomidous, I. N. Sarkar, K. Takabayashi, A. Ziegler, A. T. McCray, *et al.*, "Data analysis and data mining: current issues in biomedical informatics," *Methods of information in medicine*, vol. 50, no. 6, p. 536, 2011.

[17] D. Otasek, C. Pastrello, A. Holzinger, and I. Jurisica, *Visual Data Mining: Effective Exploration of the Biological Universe*, pp. 19–33. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[18] L. Marenco, T.-Y. Wang, G. Shepherd, P. L. Miller, and P. Nadkarni, "Qis: A framework for biomedical database federation," *Journal of the American Medical Informatics Association*, vol. 11, no. 6, pp. 523–534, 2004.

[19] V. Y. Bichutskiy, R. Colman, R. K. Brachmann, and R. H. Lathrop, "Heterogeneous biomedical database integration using a hybrid strategy: a p53 cantcer research database," *Cancer informatics*, vol. 2, p. 277, 2006.

[20] W. Sperzel, R. Abarbanel, S. Nelson, M. Erlbaum, D. Sherertz, M. Tuttle, N. Olson, and L. Fuller, "Biomedical database inter-connectivity: an experiment linking mim, genbank, and meta-1 via medline.," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, p. 190, American Medical Informatics Association, 1991.

[21] F. Aubry, S. Badaoui, H. Kaplan, and R. D. Paola, "Design and implementation of a biomedical image database (bdim)," *Medical Informatics*, vol. 13, no. 4, pp. 241–248, 1988.

[22] D. Windridge and M. Bober, *A Kernel-Based Framework for Medical Big-Data Analytics*, pp. 197–208. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[23] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining.* MIT press, 2001.

[24] J. E. Vogt, "Unsupervised structure detection in biomedical data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, pp. 753–760, July 2015.

[25] Y. Yang, S. J. Adelstein, and A. I. Kassis, "Target discovery from data mining approaches," *Drug Discovery Today*, vol. 17, no. Supplement, pp. S16 – S23, 2012. Strategic Approach to Target Identification and Validation: A Supplement to Drug Discovery Today.

[26] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro, "Human genome project," *The American journal of surgery*, vol. 165, no. 2, pp. 258–264, 1993.

[27] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, *et al.*, "The human connectome project: a data acquisition perspective," *Neuroimage*, vol. 62, no. 4, pp. 2222–2231, 2012.

[28] D. Marcus, J. Harwell, T. Olsen, M. Hodge, M. Glasser, F. Prior, M. Jenkinson, T. Laumann, S. Curtiss, and D. Van Essen, "Informatics and data mining tools and strategies for the human connectome project," *Frontiers in neuroinformatics*, vol. 5, p. 4, 2011.

[29] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[30] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, *et al.*, "Reactome: a knowledgebase of biological pathways," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D428–D432, 2005.

[31] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[32] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. Iny Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, *et al.*, "Genecards version 3: the human gene integrator," *Database*, vol. 2010, 2010.

[33] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, *et al.*, "Hmdb: the human metabolome database," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D521–D526, 2007.

[34] J. Schindelin, C. T. Rueden, M. C. Hiner, and K. W. Eliceiri, "The imagej ecosystem: An open platform for biomedical image analysis," *Molecular reproduction and development*, vol. 82, no. 7-8, pp. 518–529, 2015.

[35] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," *Bioinformatics*, vol. 18, no. 1, pp. 207–208, 2002.

[36] D. Tabas-Madrid, R. Nogales-Cadenas, and A. Pascual-Montano, "Genecodis3: a non-redundant and modular enrichment analysis tool for functional genomics," *Nucleic acids research*, vol. 40, no. W1, pp. W478–W483, 2012.

[37] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes," *Bioinformatics*, vol. 20, no. 4, pp. 578–580, 2004.

[38] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, *et al.*, "Gominer: a resource for biological interpretation of genomic and proteomic data," *Genome biology*, vol. 4, no. 4, p. R28, 2003.

[39] A. Karnovsky, T. Weymouth, T. Hull, V. G. Tarcea, G. Scardoni, C. Laudanna, M. A. Sartor, K. A. Stringer, H. Jagadish, C. Burant, *et al.*, "Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data," *Bioinformatics*, vol. 28, no. 3, pp. 373–380, 2011.

[40] J. Xia, I. V. Sinelnikov, B. Han, and D. S. Wishart, "Metaboanalyst 3.0—making metabolomics more meaningful," *Nucleic acids research*, vol. 43, no. W1, pp. W251–W257, 2015.

[41] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior research methods*, vol. 39, no. 2, pp. 175–191, 2007.

[42] D. Teodoro, R. Choquet, E. Pasche, J. Gobeill, C. Daniel, P. Ruch, and C. Lovis, "Biomedical data management: a proposal framework.," in *MIE*, pp. 175–179, Citeseer, 2009.

[43] B. Zupan, J. H. Holmes, and R. Bellazzi, "Knowledge-based data analysis and interpretation," *Artificial Intelligence in Medicine*, vol. 37, no. 3, pp. 163–165, 2006.