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 7, 2018

Contents

1 Introduction

Biomedical Engineers are known to extract useful information out of biomedical data. The biomedical data can come from many different sources: hospitals and universities, but also publicly known and commercial data. Currently a standard is missing to efficiently analyse those data sets. With the vast number of data sets that are available, such a standard in the form of a framework on data analysis would be valuable. It would speed up projects and give researchers a higher chance to reach the goal, due to improved efficiency.

First an extensive background on important topics for such a framework is discussed. Four different parts are 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 is formulated with several sub-questions for each of the four parts and a hypothesis as an answer for each of the four questions.

2 Background

Biomedical engineering can be seen as a specific part of engineering with a wide variety of topics. These topics are theoretical, non-experimental undertakings, but also state-of-the-art applications. Combining all of these different parts in one definition is not a simple task[1]. For this project, the focus is mainly on research and development, also known as knowledge discovery[2].

When a biomedical engineer begin with a project usually only a data set and a research goal are known. To achieve that goal from the data set, four different aspects influence the project's course and development:

> At first obviously the data itself is an influencer as the data restricts the project in several ways. Examples of such restrictions are multidimensionality, set size, data heterogeneity, missing feature values and population handling.
>
> A second well-known influencer is the main research goal. Since the biomedical engineer wants to achieve a certain goal, the approach outcome should match that goal for the research to be successful. Most goals are focused around either data mining, extracting relations from available data, or modelling, creating a model within data features.
>
> The steps to take from data to goal do not only need an approach, but also a tool to execute it. A third influencer therefore, is the availability of data analysis tools. The choice of a certain tool has a big impact on the project, as each one of them has its own advantages and disadvantages. The two well known tools are MATLAB and Python, however some engineers are using R, Java or C++ and there are still other possibilities.
>
> A last big influencer is biomedical knowledge. What experience the scientist already has with similar projects can 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 influences the approach, as already known outcomes do not have to be researched again.

Previous research projects on data mining called for a model how to retrieve patterns from data collections. Frameworks have been proposed, usually with a number of steps[3]. These suggested frameworks do not specifically fit the complete biomedical world for being too broad[4] or being too specific.[5] A customized framework is very beneficiary.

2.1 Biomedical Data

A big aspect of choosing how to set up the data analysis is the data itself. The amount of data in the biomedical world is growing at an enormous rate, faster than biomedical engineers can analyse.

Several additional challenges came up with this uncontrollable growth. These challenges are mainly focused around data volume, dimensionality, complexity, heterogeneity and quality[6, 7].

Scientists are tempted to collect abundant data, which makes data sets bigger than needed. Both in number of instances and features, data sets are harder to understand or analyse when more instances and features are available.[6] This volume problem usually is tackled by taking sub-populations of the complete set. Sub-sets can 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 becomes less of an issue. Volume does however become an issue when combining it with heterogeneity and quality[9, 10].

Not all data sets have a high number of instances that cause a big data volume. Sometimes there are relatively few instances, while the number of features is proportionally high[11]. Usually many of those features are not relevant enough for the research, however are still used for testing. Trying to remove features that are not important greatly helps finding relations between the others and create more knowledge about the research topic. Lowering the number of features also makes the data volume go down, making analysis 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 is that the number of samples is too low. If data is gathered for only a couple of patients, results will hardly ever be consistent. Most analyses require more samples and give biased results because of that. There are two ways of dealing with this low number of samples, the first one being generating more samples[13, 14] and the second one using the samples highly efficient[15, 16]. Most likely a combination fo both gives the best results.

Biomedical data can also be very complex. Useful results may be present, however it is very hard to obtain it. Examples of complex data are images, several biomedical signals and temporal data. Details of the useful results that are present in images is for example very hard to detect, the temporal data can vary quite much over time and the biomedical signals can be hard to combine with static biomarkers.[17] This aspect benefits from exchanging knowledge with other research areas that specialize in mining of those complex data sets[9, 18].

The biggest challenge comprises of aligning different data sets. No standard for data sets is available and therefore data sets differ greatly from each other. Data is weakly structured or even unstructured[10] and variables are processed differently due to other protocols or the collectors' preference of representation[19]. Also the variety of data is hard to combine when sources are fundamentally different. When parts of the data qre images, another part is a table from the laboratory and a third part is textual remarks of the doctor, standardizing merging those three is much harder than merging three lab sets. Those merges are also very prone to errors, as imprecisions can be vastly different between those data sets. No tool works directly with these raw data sets and preprocessing almost definitely has to occur beforehand[3, 9].

A last challenge is about data quality. The data is usually gathered by doctors and laboratory workers. Since the data is manually gathered by humans, the data has a relatively high error rate. The data can be quite noisy, values can be inconsistent, wrongly entered or even missing.[3] Not only human errors cause the data quality to drop, but the heterogeneity, as well. Two hospitals can have different protocols for the same treatment and sample different biomarkers for that protocol. Due to that difference, biomarkers can be missing for some of the entries. The time of data gathering is also a big factor as some biomarkers change greatly over time. The databases are usually also built for financial purposes and not for research, which can hurt the quality.[17]

These challenges within the data are greatly discussed.[18] Many proposals to tackle them are made, however none is actually widely adopted, yet, as a global standard for databases. Also, with the uncontrolled growth in biomedical data, it will become hard to have such a standard recognised all over the world.[19, 20, 21, 22, 23, 24]

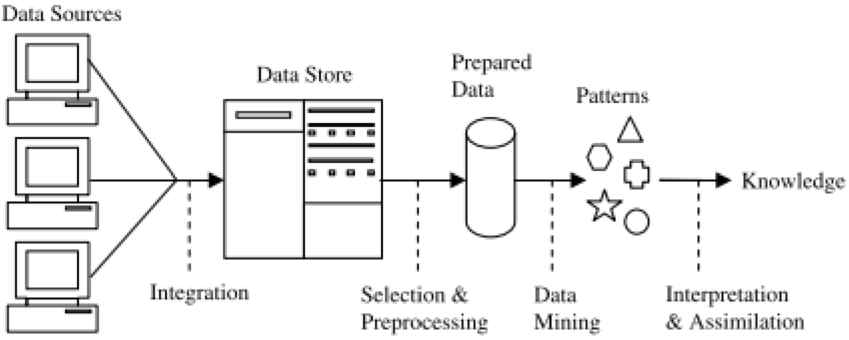

Figure 1: A schematic overview how a project involving data mining is done. Multiple data sets are integrated in one database. Feature selection and preprocessing take place to prepare the data. Then the data is mined to find patterns. These patterns are then interpreted and assimilated to discover knowledge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

2.2 Research goal: Data Mining

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

Data Mining gains multiple definitions over the years. The most adopted definition is 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[25]." As this definition states, the main goal is to find new and useful insights and patterns through datasets, that can be used in further decisions or hypotheses[6, 26]. It is one of the links that brought data and knowledge together[17]. A data mining project follows a certain layout. (Figure 2.2), which in the end leads to further knowledge that aids the medical world.[2]

Since biomedical data is a wide scope, data mining has several specialisations in different directions. An example will be text data mining, used to find related articles with websites such as Pubmed and Google Scholar. These articles can mention similar genes, diseases or proteins and give additional information about them. Another example is microarray data mining. This type of data mining focuses on extracting entities and pathways that define a disease or other phenotype. Two other data mining types are proposed. One focuses on extracting useful information out of mass spectrometry data points, called proteomic data mining. A second focuses more on a molecular level and how these molecules affect different cell types[27].

Whichever splits are made in the data mining aspect for different areas in biomedical data analysis, the techniques are mostly based on statistical analysis and machine learning. A discussion of these two types have 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 has to be used that facilitated the use of those algorithms. Several basic tools are available and widely used[1]. Since biomedical data has specific characteristics, separate tools or frameworks consisting of those tools are made. These are 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 are connected as well, for optimized use.

[1] Biomedical Data Analysis - *Tim Beishuizen*

To store biomedical data several projects are started. Famous project examples are the Human Genome Project (HGP) that tries to sequence the entire human genome[28] or the Human Connectome Project (HCP) that focuses on the brain connectivity[29]. Both consist of an enormous amount of data that needs to be stored and for both a tool is created for use in these project, for instance the ConnectomeDB for the HCP[30]. Similar projects resulted in biomedical databases, such as KEGG[31], Reactome[32], NCBI[33], GeneCards[34] and HMDB[35]. These databases store data about genes, metabolites and pathways and are used by numerous scientists to completely understand those. More databases are made or are in development, all of them for a specific domain in biomedical research.

One immediate issue with all of theses existing databases is that hardly any framework or tool can be found for synchronisation of these databases. Their use is limited to a specific area only without any possibilities to switch in between. An all encompassing framework is not made. For that reason, multiple scientists make a tool specifically for their specific area. New scientists have a hard time using the newly created tool and try to create a new one themselves. To stop that cycle a new framework that combines aspects of the available tools will be very helpful[44].

Several examples of tools and frameworks of biomedical data mining are available. A framework designed specifically for biomedical data was ImageJ[36]. ImageJ can be used for biomedical image analysis and consists of image processing and analysis techniques. Another example is Genesis[37], a platform for gene expression analysis based on Java and consisting of various preprocessing and clustering algorithms. Similarly the tools GeneCodis3[38], FatiGO[39] and [40] are created to interpret results from genomics generated data. Similar tools are made for metabolomics, such as Metscape 2[41] and MetaboAnalyst[42]. A last example is the tool G*power3[43] that focuses more on statistical analysis of biomedical data.

2.4 Biomedical Knowledge

A last important aspect for data analysis is the available knowledge on the topic. Numerous scientists have done research on various biomedical topics. The knowledge gathered from these topics is available in several different ways[18]. Usually this information is found in a textual manner, in books, papers and articles describing the project results. Other information is found in the earlier mentioned databases, tools and designed algorithms. A third type of knowledge is implicit, scientists are expected to have certain skills, or should be trained for if not[45]. All of these are interesting in their own ways.

Literature knowledge is the basis for all research. The knowledge found on several different research areas can result in new insights and ideas for unexplored parts. The projects generated this way become knowledge-driven and have hypotheses based on gathered information[18]. Not only at the start of research projects knowledge is used, as it is a guide throughout the whole process. In the final project steps the results should be used to improve or extend the existing knowledge, so it can be used in new projects[45].

When looking at projects on implementation level, more interesting data knowledge can be gathered from databases, tools and algorithms. These three can be used to test and answer hypotheses derived from textual knowledge. Mainly these are used for data analysis as known data is gathered from databases, tools are used for their analysis properties and algorithms are 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[18].

A third knowledge type is the implicit knowledge of the scientists conducting a research project. These scientist can 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 are made with a certain type of scientist in mind as well, making it important that the desired type of scientist is able to use it[45].

These types of knowledge can cause several issues. Projects will not be conducted properly if literature does not back up initial points of view and assumptions. On the other hand data knowledge that stems from other research projects can 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 is to create a framework for more efficient data analysis. Based around the four main influencers of data analysis, a main research question is made and divided in five separate sub-questions. The main question is:

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

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 extension required for completion of the framework?
>
> *Knowledge driven:* How must both availability- and lack of knowledge be included in the framework?

3.1 Hypotheses

Since there are five different sub-questions, also five hypotheses are formed. These hypotheses define the layout of the framework.

> *Data driven:* The available data mainly influences the choice for a certain data analysis approach when preprocessing. Several preprocessing actions should usually be done before actual research is possible.
>
> *Target driven:* The research goal is mainly important for the data mining part of the project. This means that multiple steps of data mining are taken in an increasing zoomed in level to eventually achieve the final goal.
>
> *Tool availability driven:* Three tools were chosen to be good additions to the framework: SciPy, scikit-learn and TPOT.
>
> *Tool extension driven:* Three main extensions are thought to be beneficial for the framework: a global analysis tool, an extension for TPOT on a preprocessing level and a database integration tool.
>
> *Knowledge driven:* As biomedical engineers are the main target fo the group, the framework will be directed mainly to them, with additions such as a GUI and useable data sets as examples for how to use the framework.

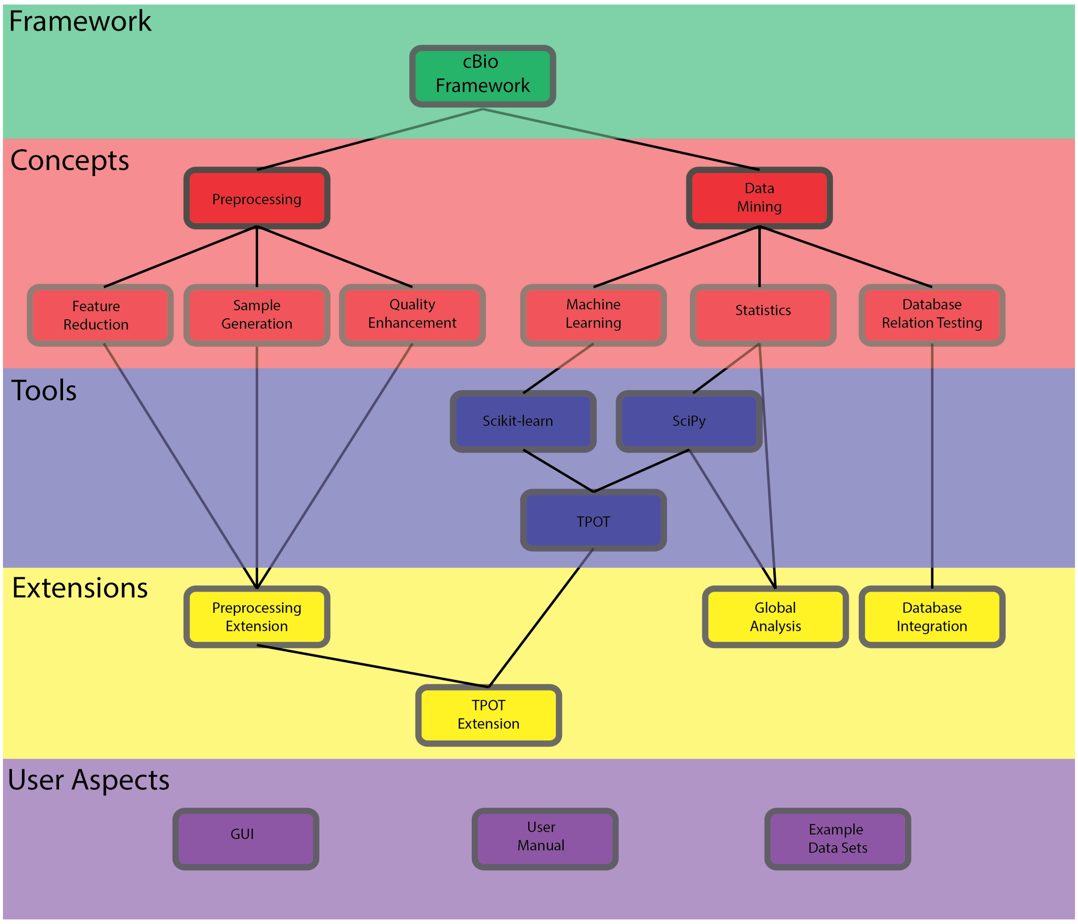

Figure 2: A schematic overview of how the framework eventually should look like. Every colour implied a different layer of implementation and corresponds to different approaches: (red) data driven and target driven, (blue) tool availability driven, (yellow) extension driven and knowledge driven.

4 Implementation Approach

The framework will be created with at the base these five questions. For each question a different approach is given how to handle that in the framework. For illustration of the framework a tree is created (Figure 2) With help of this tree every approach is shown within the complete framework.

4.1 Data Driven Approach

A major part in a data driven analysis approach is the presence of challenges with preprocessing the data. Data volume, dimensionality, complexity, heterogeneity and quality challenges all are issues to be tackled, some of them more interesting than others. The most approachable three challenges are 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 can be tackled by various feature selection and reduction algorithms, such as principal component analysis, significance testing and clustering. Low data volume can 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. These three preprocessing aspects all are separately addressed in the framework at the concepts section (Figure ??).

4.2 Target Driven Approach

To reach a certain goal, several steps have to be taken. While the steps usually are vastly different, it always starts with a global analysis to localize possible issues and interesting features in the data. A statistical analysis is very beneficial for doing such a global analysis. Adding an easy possibility for such a global analysis would benefit the user.

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

The final steps to achieve the goal is very personalized for every project. To approach these final steps earlier results are used to obtain a final idea. This means that the scientist itself has the best idea how to achieve that goal with the knowledge gathered from the first couple of steps. Therefore it is not useful to guide someone in this part of the data analysis and no special focus is needed for this step.

A small summary: he global analysis and basic techniques are part of data mining. Three parts are addressed for data mining: machine learning, statistics and database relation testing. These three data mining aspects all are separately addressed in the framework at the concepts section (Figure ??).

4.3 Tool Availability Approach

Tools to be included in the final framework must be focused around the data driven and target driven desired techniques. This means that it should include tools for feature reduction and selection, sample 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 provides already several algorithms for this problem and may be a good start for the framework. Several feature selection and machine learning algorithms are present and easily useable. A very welcome addition to scikit-learn is the Tree-based Pipeline Optimization Tool (TPOT) that automatedfinding the best machine learning algorithm with the best hyperparameters. TPOT uses both algorithms from scikit-learn and algorithms made by the user. These two are good tools for a successful data mining project.

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

To summarize, the three tools that will be a big part of the framework are scikit-learn, SciPy and TPOT. These three are addressed at the tools section (Figure ??).

4.4 Tool Extension Approach

With the mention of Scikit-learn, TPOT and SciPy, most aspects desired in the framework are at least partly covered. These three are 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 is needed to TPOT more suitable for biomedical data.

With the mentioned tools a good base for most of the desired parts in the framework are mentioned. The missing aspects that are not discussed, yet, would be the three types of preprocessing. These three and possible more algorithms for the other aspects can be programmed to show more variety in preprocessing and data mining. Methods that are created for these extensions are then also to TPOT, for more automated preprocessing.

Some statistical methods are available for analysis in SciPy. These statistical methods must be called separate from each other in SciPy and no selection is made for an initial global analysis. Therefore a new method for doing that global analysis can be made, that borrows methods from SciPy to give the use insight in the dataset. This means it needs basic statistical methods, such as descriptive statistics computation, significance testing and multicollinearity testing.

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

The four different extensions are: preprocessing extension, TPOT extension, global analysis and database integration. All four are added to the framework at the extension section (Figure **??**).

4.5 Knowledge Driven Approach

The framework is created for Biomedical Engineers. As Biomedical Engineers, the Computational Biology group of the Biomedical Engineering Department of the University of Technology Eindhoven are taken as a reference group. Since this group mainly consists of either experienced employees that are already doing a research projects or less experienced students, the best focus for this framework is biomedical students. They benefit the most from such a framework that can help them at the start of their projects. The scope of the framework is then people that have some minor programming skills and need guidance for better understanding and efficiency.

Since inexperienced programmers do not know immediately how to program in possibly a different language, a graphical user interface (GUI) is helpful, as well. This GUI can mainly help in the global analysis, but maybe also in running some initial calculations with TPOT. When using this GUI, programming is less important as a skill and therefore people do not need to figure out how the framework is designed. For programmers that need to use more specific methods, it still is possible to use it as packages for their own projects.

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

The GUI, user manual and exemplary data set handling are the three ways to approach the knowledge driven aspects. These three are added to the framework at the Usability section (Figure **??**).

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] D. Dunbar and G. Humphreys, "A spatial data structure for fast poisson-disk sample generation," in *ACM Transactions on Graphics (TOG)*, vol. 25, pp. 503–508, ACM, 2006.

[14] L. Devroye, "Sample-based non-uniform random variate generation," in *Proceedings of the 18th conference on Winter simulation*, pp. 260–265, ACM, 1986.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.