Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.2622

ACDBio: The Biological Data Computational Analysis group at ICMC/USP, IFSP, and Barretos Cancer Hospital

Adenilso Simao [**University of São Paulo** | ***adenilso@icmc.usp.br***]
Adriane Feijó Evangelista [**Barretos Cancer Hospital and Oswaldo Cruz Foundation** | ***afevangelista@alumni.usp.br***]
Alfredo Guilherme Souza [**University of São Paulo** | ***alfredo@usp.br***]
Cynthia de Oliveira Lage Ferreira [**University of São Paulo** | ***cynthia@icmc.usp.br***]
Jorge Francisco Cutigi [**Federal Institute of São Paulo** | ***cutigi@ifsp.edu.br***]
Paulo Henrique Ribeiro [**Federal Institute of São Paulo** | ***phribeiro@ifsp.edu.br***]
Rodrigo Henrique Ramos [**Federal Institute of São Paulo** | ***ramos@ifsp.edu.br***]

University of São Paulo, Av. Trabalhador São-carlense, 400, São Carlos, SP, 13566-590, Brazil.

Received: 2 April 2022 • **Published:** 17 February 2024

Abstract Recent advances in biological and health technology have resulted in vast digital data. However, the major challenge is interpreting such data to find valuable knowledge. For this, using computing is essential and mandatory since quick data processing and analysis, allied with knowledge extraction techniques, enable working effectively with large biological datasets. In this context, the ACDBio group works with the computational analysis of biological data from different sources, aiming to find new information and knowledge in data or answer questions that are not yet known. So far, the group has worked on several challenging topics, such as identifying significant genes for cancer topological analysis of genes in interaction networks, among others. The group uses computational techniques such as complex networks and their algorithms, machine learning, and topological data analysis. This article aims to present the ACDBio group, and the main research topics worked on by its members. We also present the main results and future work expected by the group.

Keywords: Bioinformatics, Biological Data Analysis, Computer Science

1 Introduction

A considerable amount of data has been generated in many fields in recent years. In biology, especially in genomics, data originate after the emergence of next-generation sequencing. Clinical data became available due to the increasing deployment of information systems for health management. There is the challenge of analyzing and interpreting such data to obtain valuable information. Thus, the ACDBio group emerged, whose main objective is to study biological data in general, especially data related to cancer, epidemiological diseases, and network data.

ACDBio is a research group that started its activities in late 2015, where two members studied a classic problem in cancer genomics: the identification of significant genes for cancer through computational methods. In 2017, the group started to count on a member who is an expert in Cancer Genomics. Currently, the group has nine people, including professors, graduate students, and undergraduate students. Furthermore, it is an inter-institutional group with researchers from three institutions: University of Sao Paulo (ICMC/USP), Federal Institute of Sao Paulo (IFSP), and Barretos Cancer Hospital (HCB). The group still maintains partnerships with Federal University of Rio de Janeiro (UFRJ). The group has been working on several research topics: cancer genomics, dynamics of epidemics, topology and resilience of complex biological networks, and topological analysis of biological data. Figure1 presents an overview of the research topics and the main computational approaches used in investigating these topics.

We have developed studies with relevant results. The main contributions so far are: Identification of cancer driver genes, using some types of data (e.g., gene interaction network and mutation data) and cancer genomics pattern (e.g., mutual exclusivity) and how different types of mutation can impact and influence genes that are in the network neighborhood; Detection of false-positive driver genes through the induction of supervised machine learning models that use as features mutation data and centrality measures of genes in interaction networks; Modeling Reactome's Super Pathways as networks and discovering topological distinction between cancer driver genes and non-cancer driver genes that can be explored to find new driver genes.

This paper is organized as follows: Section 2 presents the main research topics in which ACDBio is involved. The published works are described with the main obtained results. Section 3 describes the main datasets used in group researches and analyses. Next, Section 4 shows research topics that ACDBio is currently starting the research, or it is interested in working in the future. Finally, Section 5 presents the final considerations about ACDBio and this work.

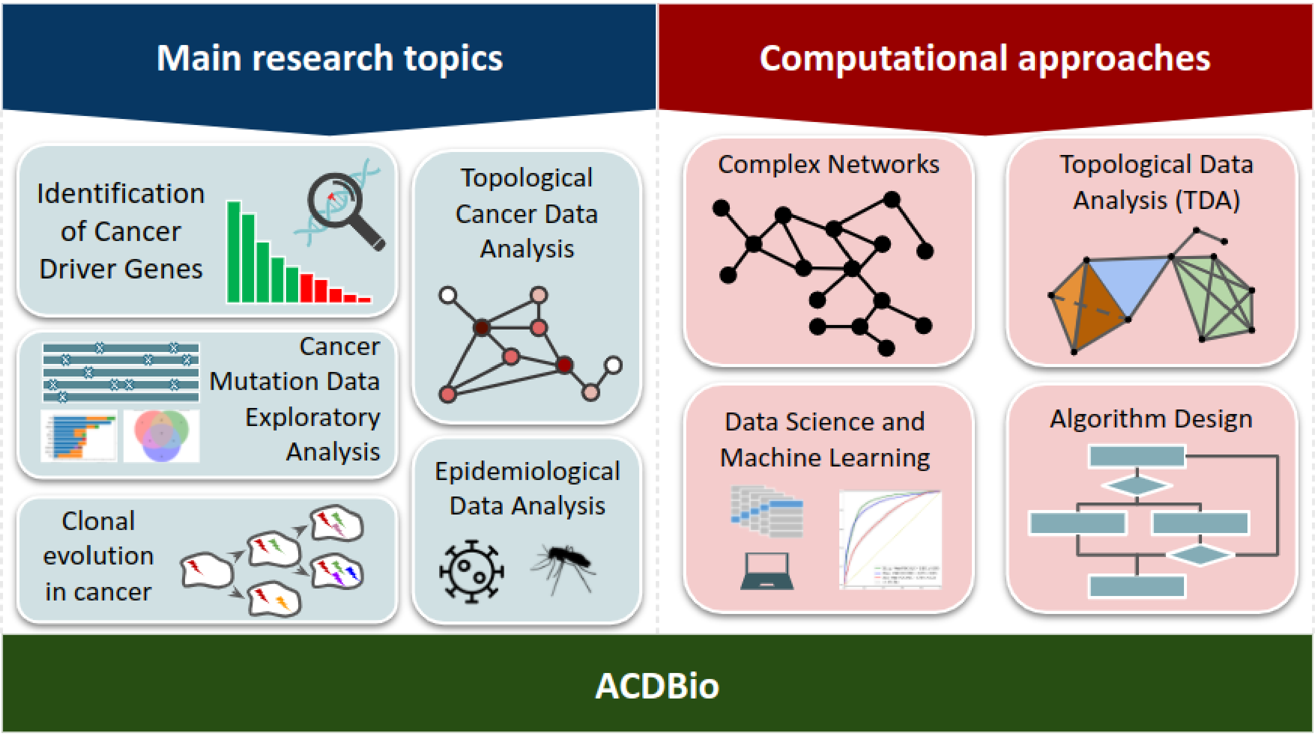
Main research topics
Computational approaches
Identification of Cancer Driver Genes
Topological Cancer Data Analysis
Cancer Mutation Data Exploratory Analysis
Clonal evolution in cancer
Epidemiological Data Analysis
Complex Networks
Topological Data Analysis (TDA)
Data Science and Machine Learning
Algorithm Design
ACDBio

Figure 1. Group overview: In the left side is presented the main research topics covered by the group, while in the right side in presented the main computational approaches used to work in the topics.

2 Research topics

In this section, we present the main research topics of ACDBio. We describe concluded works with results, the published article, and ongoing works. Overall, the studies address cancer genes and other biological data analyses. The works associated with cancer genes cover the discovery of cancer driver genes using genes networks, an overview of the computational methods for cancer genes, explore the topological characterization of cancer genes in pathways networks, and the discovery of false-positive cancer genes. The other works explore the differences in cancer datasets, the Topological Data Analysis (TDA), and its application to biological data analysis. Lastly, we present epidemiological data analysis concerning COVID-19 and Dengue.

2.1 Cancer genes identification

Cancer is caused by the accumulation of genetic alterations during an individual's life. Such alterations, called genetic mutations, can lead the cell to disordered and uncontrolled growth, which can cause cancer. New genome-sequencing technologies, called Next-Generation Sequencing (NGS), enable fast and cost-effective genomic sequencing. As a result, a large volume of biological data, including cancer data, can be processed and analyzed to find useful clinical information. In this context, Cancer Bioinformatics develops computational methods to interpret cancer data and help with cancer understanding. One of the categories of computational methods includes those that lead to the problem of identifying significant mutations (driver mutations) and their associated genes (driver genes) for cancer development. This is a challenging problem once most of the mutations are not related to cancer. In this context, the computational methods employ several strategies to deal with this problem, e.g., gene interaction networks.

In the following subsections is presented the main results achieved by ACDBio on cancer genes identification research topic.

2.1.1 Discovering cancer genes through gene network, weighted mutations and mutual exclusivity pattern

This work describes a computational method for prioritizing related genes significantly mutated in cancer. First, the seminal idea of the method was published as short paper at SBCAS (Simpósio Brasileiro de Computação Aplicada a Saúde) [Cutigi *et al.*, 2019b]. After that, an extension of this previous publication was submitted and accept for publication and presentation at BSB (Brazilian Symposium in Bioinformatics) [Cutigi *et al.*, 2019a]. This work presents a flexible computational method named GeNWeMME (Gene Network + Weighted Mutations + Mutual Exclusivity). Such a method has four steps, which is illustrated in Figure 2 (extracted from [Cutigi *et al.*, 2019a]).

The method uses mutation data, gene interaction networks and mutual exclusivity patterns prioritizing groups of significant genes in cancer. All these aspects are used according to the objective of the analysis by cancer genomics professionals, that can choose weights for each aspect. The prioritized related genes are mutated in most patients and present a pattern of mutual exclusivity. Experimental validation was conducted using four types of cancer. Such validation showed that it was possible to identify known cancer genes and suggest others for further biological validation. Also,

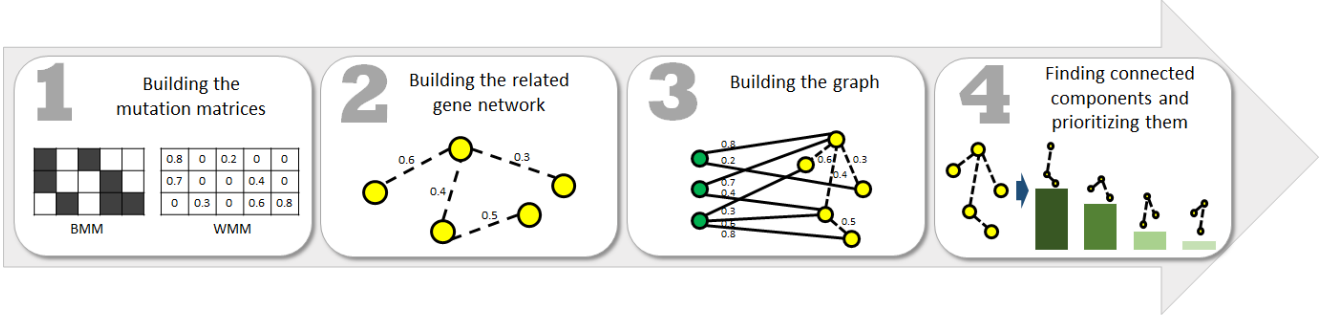

Figure 2. An overview of the GeNWeMME method.

GeNWeMME could prioritize genes with a low frequency of mutation.

2.1.2 Computational perspective of computational methods

Many computational methods to discover significant genes for cancer have been developed. Such computational methods are complex and their understanding is most of time difficult, especially for research from Computer Science. For this, a tutorial was produced with details about some classical computational methods, from a computational perspective, with the transcription in an algorithmic format for easy access by researchers. This work was published at the *Journal of Bioinformatics and Computational Biology (JBCB)* [Cutigi *et al.*, 2020a].

The paper describes some classical computational methods that identify driver mutations from a computational perspective. It focuses on algorithms that use gene networks and mutual exclusivity patterns, in which the computational and biological aspects of the methods are discussed and compared. The methods were transcribed in an algorithmic format to facilitate their comparisons and understanding, focusing on showing them from a computational perspective. The paper briefly describes some other related works, thus summarizing such methods. It also discusses their computational complexity and how they can be evaluated and compared. With this work, the expectation is to broaden researchers' understanding of the computational aspects of important and classical methods for identifying significant mutations in cancer.

2.1.3 Detecting possible false cancer genes

Although computational methods have been used to identify significant genes for cancer, they can misclassify some genes as significant, thus requiring expert curation to filter their findings [Bailey *et al.*, 2018]. Such misclassification is due to some genes (referred to as false-positive-drivers or false-drivers) exhibiting characteristics of being significant for cancer, despite not being involved in its initiation and progression. To avoid misclassifying false drivers as drivers, a computational method was proposed to classify possible driver genes as real or false-driver.

The method was described in a paper published and presented at BSB (Brazilian Symposium in Bioinformatics) [Cutigi *et al.*, 2020b]. The classification of genes in possible false drivers is performed through a supervised machine-learning approach. Random Forest (RF) and Support Vector Machine (SVM) models were induced by features extracted from mutation data and gene network interactions. Both models were evaluated using machine learning classical metrics, and they achieved satisfactory classification performance, benefiting from the combination of mutation and gene interaction features. Figure 3 (extracted from [Cutigi *et al.*, 2020b]) shows the supervised machine learning process to induce the classification models, followed by their evaluation.

2.1.4 Topological characterization of cancer driver genes using Reactome super pathways networks

The study of topological characteristics of genes in the network and pathways is an important topic once it can contribute to understanding the role of drivers and their genes in the networks. The work of [Cutigi *et al.*, 2020b] shows that gene network centrality measures increase the potential of detecting possible drivers and false drivers. Furthermore, a great number of network-based methods use information about networks to identify significant genes in cancer [Ozturk *et al.*, 2018]. In this context, the ACDBio group published in the proceedings of the BSB (Brazilian Symposium in Bioinformatics) in 2021 a study on a topological characterization of cancer driver genes using Reactome super pathways networks [Ramos *et al.*, 2021]. The study models super pathways as complex networks to observe the topological characteristics of driver genes and their central role in such networks, aiming to investigate the hypothesis that driver genes are topologically different from other genes in the same pathway.

The paper shows significant differences in some centrality measures between drivers and non-drivers. For example, the measures of betweenness and closeness play an essential role in characterizing the drivers. Also, concerning the resilience of super pathways networks, drivers can help to understand the impact of mutations in biological functions and their influence on cancer. Considering Programmed Cell Death Pathway, we observe the drivers' central role in maintaining the network's topological integrity. In other networks, drivers' behavior was similar to the random removal of nodes. At the same time, some networks show remarkable resilience even to hub attacks. These results show that super pathways networks have distinct topologies and particular roles for drivers. Groups of pathways that share similar results in centrality measures differ in the resilience of inten-

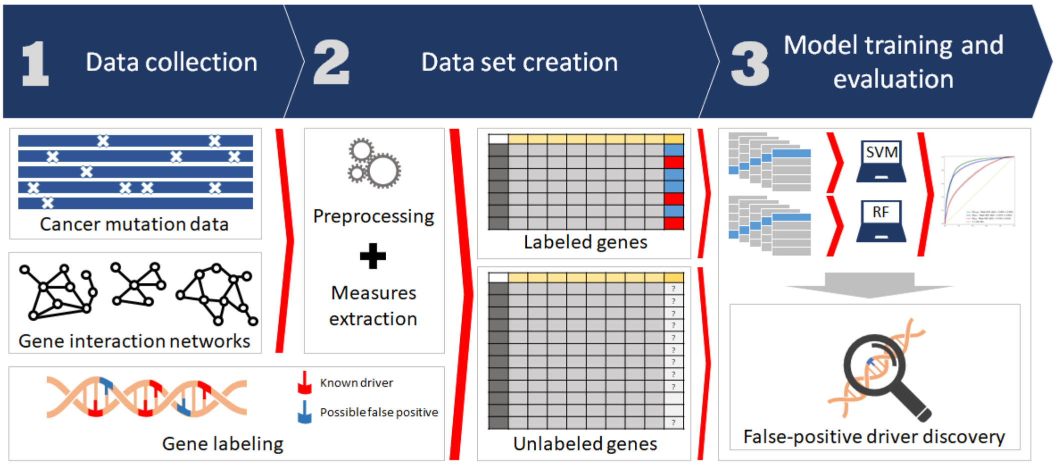

Figure 3. An overview of the machine learning-based approach to detect false drivers.

tional attacks. These findings reinforce the need to diversify the analysis of driver's topology. Also, treating each super pathway as an individual system may provide more reliable results.

2.1.5 Discovering cancer genes using weighted mutation and asymmetric spreading strength in networks

Recently, a new computational method, called DiSCaGe (**Di**scovering **S**ignificant **Ca**ncer **Ge**nes), was proposed by the group and published at the *Nature Scientific Reports* [Cutigi *et al.*, 2021]. Figure 4 (extracted from [Cutigi *et al.*, 2021]) presents DiSCaGe overview.

DiSCaGe takes advantage of weighted mutation frequency and network neighbors' influence. The weighted mutation frequency considers the mutations and the possible functional impact for cell carcinogenesis. The network influence is through an asymmetric spreading strength measure that can quantify how a mutation can affect the network neighborhood. DiSCaGe was evaluated in six cancer types using their mutation datasets and two gene interaction networks. The results showed DiSCaGe's potential for discovering known cancer-related genes, including genes with low mutation frequency, and cited in research papers as cancer-related genes. Furthermore, DiSCaGe also suggests possible novel cancer genes.

2.1.6 Investigation of the performance of driver mutation identification methods using biological networks and enriched biological networks

Several computational methods allow identifying cancer-related genes (driver mutation) through patient mutation data and biological networks. Usually, networks are not built focusing on cancer-related biological activities because they are designed for general use. In this study, we investigate the performance of methods for identifying driver mutations using biological networks and enriched biological networks, applying a gene prioritization method to classify genes associated with cancer understudy in the biological network. We selected six types of cancer to be used as a case study, such as: Bladder Cancer (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma (GBM), Pancreatic Adenocarcinoma (PAAD), Prostate Adenocarcinoma (PRAD), and Stomach Adenocarcinoma (STAD). The results indicated that the enrichment method helped identify different driver genes in all cases. This study was recently submitted to the SBCAS (XXII Simpósio Brasileiro de Computação Aplicada à Saúde) and is currently under review.

2.2 Biological data analysis

With the development of the ACDBio group, we branched our work in areas beyond cancer genes. A study about cancer datasets served as a foundation for understanding the input used in many computational methods exploring the disease and how different datasets from the same type of cancer can potentially modify the methods' output. Topological Data Analysis (TDA) is a recent field that extracts information about the data structure and has been used to explore biological data. With the increased data about COVID-19, we investigate how vaccines impact the survival rate among hospitalized and ICU Brazilian patients. A partnership with Federal University of Rio de Janeiro motivated the creation of a study of Dengue contagious dynamics with temporal data in Rio de Janeiro.

2.2.1 Cancer mutation datasets analysis

There is a huge variety of mutation data in public databases. However, it is not feasible to use all available data in every analysis; thus, a data subset must be selected. Considering this context, the ACDBio group published in 2020 [Ramos *et al.*, 2020] in the proceedings of the SBCAS (Simpósio Brasileiro de Computação Aplicada à Saúde), a work whose objective was to investigate and understand the mutational characteristics presented in different cancer mutation datasets of the same type of cancer. To achieve this goal, exploration and visualization of cancer mutation data were performed. Several analyses were presented for three common types of cancer: 1) Breast Invasive Carcinoma (BRCA); 2) Lung Adenocarcinoma (LUAD); and 3) Prostate Adenocarcinoma (PRAD). For each cancer type, three distinct datasets were analyzed to understand whether they have significant differences or similarities. The analyses showed that BRCA

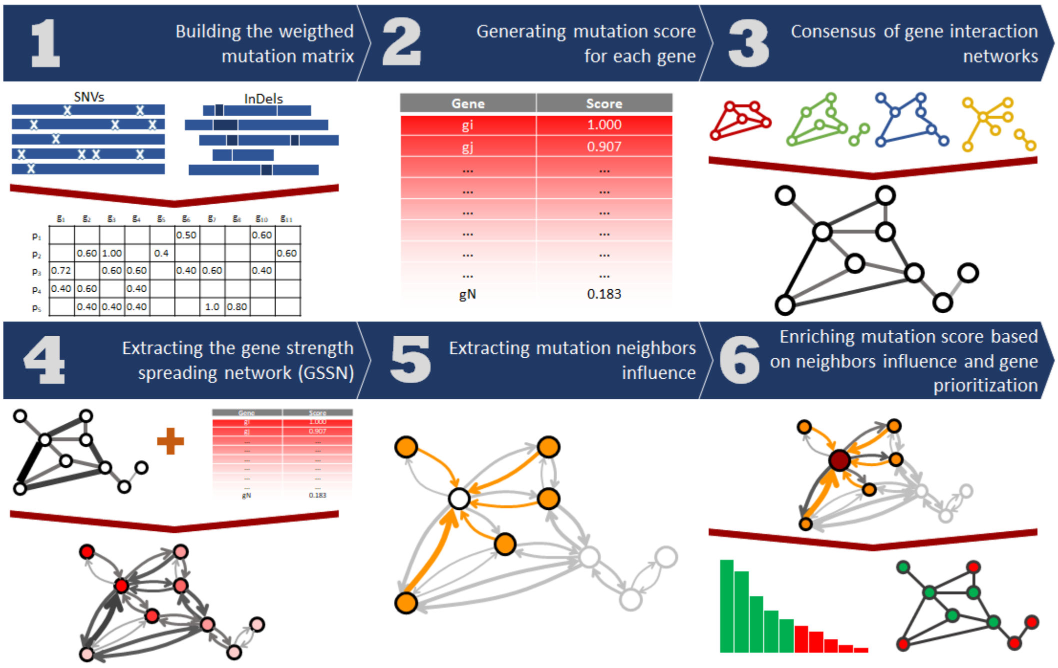
1 Building the weigthed mutation matrix
2 Generating mutation score for each gene
3 Consensus of gene interaction networks
SNVs
InDels
Gene
Score
gi
1.000
gj
0.907
gN
0.183
4 Extracting the gene strength spreading network (GSSN)
5 Extracting mutation neighbors influence
6 Enriching mutation score based on neighbors influence and gene prioritization

Figure 4. An overview of DiSCaGe method.

and LUAD have evidence of similarity among their datasets, while PRAD is likely heterogeneous.

2.2.2 Topological data analysis

Another interest in the ACDBio group concerns topological data analysis (TDA) tools. TDA is a recent field that emerged from the work of [Edelsbrunner *et al.*, 2000] on persistent homology and was popularized in a reference article by [Carlsson, 2009]. TDA is motivated primarily because topology and geometry provide an approach to inferring qualitative and sometimes quantitative information about the data structure [Chazal, 2016]. In that regard, TDA aims to develop mathematical and statistical tools and algorithms to infer, analyze, and explore the complex topology of data and its geometric structures. It has been used to reveal hidden subgroups of cancer patients, construct organizational maps of brain activity and classify abnormal patterns in medical images. Some applications of TDA in biological data can be seen in [Rabadan and Blumberg, 2019]. The ACDBio group has been working with TDA in the context of cancer data, such as studying the topological characteristics of disease-associated pathways. Other applications of interest are generally related to biological data, such as COVID-19 or Dengue virus data, to find patterns hidden behind the point cloud.

2.2.3 The survival rate among unvaccinated, first dose, and second dose brazilian hospitalized and ICU COVID patients by age group

The ACDBio group explored the lethality among hospitalized COVID-19 patients with one dose, two doses, and unvaccinated. We used a Brazilian nationwide surveillance repository of severe acute respiratory disease for hospitalized patients with data for 1,177,151 hospitalized and ICU COVID-19 patients in 2021. The data was preprocessed and divided into age groups between hospitalized and ICU patients. The results showed statistical evidence for hospitalized patients that lethality increases with age and decreases with vaccines, especially with the second dose. We also point out a significant difference among age groups and between hospitalized and ICU, indicating the need to separate these groups when analyzing lethality and comorbidities. Furthermore, we also explore the dynamics of the symptoms over time. This study was recently submitted to the SBCAS (XXII Simpósio Brasileiro de Computação Aplicada à Saúde) and is currently under review.

3 Datasets

The ACDBio works with different and complementary datasets. In this section, we present an overview of the most used datasets.

3.1 Molecular datasets

With the advancements in NGS technologies, many databases make available data concerning cellular function and cancer mutation. This data makes possible the study of complex diseases through computational approaches. Here we present three types of molecular datasets that ACDBio often uses.

3.1.1 Mutation Annotation Format

Cancer studies, such as [Ciriello *et al.*, 2015], made available many data files. One of these files is the Mutation Annotation Format (MAF), a tab-delimited file containing mutations found in samples. Each patient in the study has one or more samples, and each sample has one or more genes with one or more alterations (mutations).

The Genomic Data Commons, from the National Cancer Institute (NCI), defines the pipeline used to create the MAF and the 126 columns found in the file[1]. Many of these columns store duplicate information that used to be indexed by different systems and databases, as well as some metadata fields. Of the 126 fields present in the file, nine are frequently used in computational methods:

- Hugo Symbol: Gene symbol following the Human Genome Organisation (HUGO) standards.
- Chromosome: The affected chromosome.
- Start Position: The mutation start to coordinate.
- End Position: The mutation end coordinate.
- Reference Allele: The strand reference allele includes the deleted sequence for a deletion or ”-” for an insertion.
- Tumor Seq Allele: Primary data genotype for tumor sequencing (discovery) allele
- Variant Classification: The translational effect of variant allele, example: Missense, Silent, frameshift deletion.
- Variant Type: The mutation type, for example, TNP (tri-nucleotide polymorphism), DNP (di-nucleotide polymorphism), ONP (oligo-nucleotide polymorphism).
- Tumor Sample Barcode: The barcode for the tumor sample. It is a unique identifier for the sample and the patient.

3.1.2 Protein networks

A gene establishes complex interactions with other genes and their produced proteins. These interaction can be modeled as networks, a natural way to represent complex biological systems [Kim *et al*., 2016]. Protein interaction networks are used mainly in Cancer Genomics. In this case, proteins are nodes, and edges connect proteins that interact in some aspect.

Many databases are sources of information about networks, e.g., Human Protein Reference Database (HPRD) [Peri *et al*., 2003; Keshava Prasad *et al*., 2009], High-quality INTeractomes (HINT) [Das and Yu, 2012], Reactome Functional Interactions (ReactomeFI) [Jassal *et al*., 2020], and Human Reference Interactome (HuRI) [Luck *et al*., 2020].

3.1.3 Pathways

Pathways are sets of genes that interact and are responsible for the emergence of specific biological functions. Each pathway works as building blocks of a cell’s complex system. Some examples of pathways are: Circadian Clock, DNA Repair, and Programmed Cell Death [Jassal *et al*., 2020].

The pathways are considerably smaller than the protein networks and are associated with specific molecular functions. Using pathways in computational methods decreases the complexity and increases the significance compared to the whole protein network. While this approach has advanced, it lacks the generality found when using protein networks. The ACDBio group aim to combine pathways and protein network in its methods.

3.2 COVID dataset

There is a worldwide interest in studying COVID-19 data, with led to the creation of public databases. Brazil has a database called SIVEP-Gripe, a nationwide surveillance repository of severe acute respiratory disease for hospitalized patients. Considering only 2021, the SIVEP-Gripe store data for 1,177,151 hospitalized and ICU COVID-19 patients.

For each hospitalized patient, the dataset covers personal and regional information (e.g., age, gender, race, city, state, and hospital name). It also presents information about symptoms (e.g., fever, cough, saturation) and comorbidities (e.g., heart disease, pulmonary disease, asthma, kidney disease). There is also temporal information like date evolution (dead or alive), date hospitalization, date of first symptoms, and date of vaccine doses.

4 Future directions

The studies presented in the last section opened opportunities for new researches the ACDBio group will develop in the second half of 2022. This section presents two future works related to cancer that will deepen the understanding of cancer genes: clonal expansion and mutual exclusivity. This section also presents future work regarding the spread of Dengue contamination over more than ten years in Rio de Janeiro state. This study will be executed in partnership with the Universidade Federal do Rio de Janeiro since they have the data and specialized researchers in epidemiology.

4.1 Analysis of clonal evolution in cancer

Driver mutations provide the cell in which they occurred with an evolutionary advantage over other cells, making it better suited to its local micro-environment, enabling it to proliferate quicker than other cells and generate more daughter cells. This process is called “clonal expansion” [Greaves and Carlo, 2012; Vogelstein *et al*., 2013; Dujon *et al*., 2021]. Most cancers evolve from a single cell with driver mutations through a series of clonal expansions.

As the tumor develops, the cells continue to acquire more driver and passenger mutations. A tumor cell that acquires an additional driver mutation that causes clonal expansion will generate a subpopulation of cells with mutations that are not present in all cells in the tumor. This population of subclonal cells can be identified through a set of shared mutations. Clonal expansions can lead to several coexisting subclones sharing subsets of mutations. Cancers are increasingly recognized as mixtures of competing subclones [Nik-Zainal *et al*., 2012].

In this context, there are studies on the clonal evolution of cancer to infer the subclonal composition of a tumor by identifying cell populations with shared mutations. The application of subclonal reconstruction methods provides important information about tumor evolution, identifying subclonal driver mutations, patterns of parallel evolution, and differences in mutation signatures between cell populations, and characterizing the mechanisms of therapy resistance, propagation, and metastases [Dentro *et al*., 2017].

[1] https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Future directions include analyzing the main methods of clonal evolution analysis from a computational and biological perspective to identify improvements in existing methods and, eventually, propose a new method to infer the structure of the clonal population.

4.2 Mutual exclusivity

Driver mutation in pathways responsible for cell proliferation and survival is frequently associated with cancer development. It is intuitive to think that the more mutations a tumor has, the faster it progresses. However, large-scale genomics studies show otherwise: driver oncogenes often are mutually exclusive [Cisowski and Bergo, 2017]. Albeit this phenomenon is not entirely understood, recent studies point out that mutual exclusivity may be associated with tumor type and interactions between drivers' genes. A review paper on 21 computational methods for mutual exclusivity addresses key features that can cause false-positive discovery: cancer sub-type, intra-tumor heterogeneity, and imbalance of mutual exclusivity [Deng *et al.*, 2019]. Most of the reviewed methods do not consider any of these key features and thus are prone to false-positive discovery.

One hypothesis for mutual exclusivity is Functional Redundancy [Deng *et al.*, 2019]. This hypothesis is based on pathway topology and the downstream effect, where a mutation in one gene on the stream is enough to corrupt the entire pathway. Following this hypothesis, identifying mutated genes that corrupted the pathway can help understand which biological function corruption leads to cancer.

The ACDBio group will work on the hypothesis that a computational method to find mutual exclusivity, which considers qualitative and quantitative data of patients and tumors, can generate more significant results with fewer false positives.

4.3 Topological analysis on dengue contamination in the state of Rio de Janeiro

The ACDBio is starting a partnership with the Universidade Federal do Rio de Janeiro, collaborating with a biomathematical group. We will analyze temporal data for more than ten years about Dengue proliferation concerning aspects beyond heat and humidity, adding information about state subregions, socioeconomics and transport.

5 Conclusion

This paper describes ACDBio, an inter-institutional research group that studies biological data in general through computational analysis. Since its creation, the group has investigated various related topics, such as cancer genomics, epidemiological diseases, and network data. In these topics, some interesting results were achieved that were described in papers published in conferences and journals.

In future work, it is expected the study new research topics, such as clonal evolution, and mutual exclusivity in cancer, among others. Additionally, new collaborations with other research groups could be achieved, in order to expand the interdisciplinarity of the group.

References

Bailey, M. H. *et al.* (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371 – 385.e18. DOI: https://doi.org/10.1016/j.cell.2018.02.060.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Chazal, F. (2016). High-dimensional topological data analysis.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., *et al.* (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Cisowski, J. and Bergo, M. O. (2017). What makes oncogenes mutually exclusive? *Small GTPases*, 8(3):187–192.

Cutigi, J. F., Evangelista, A. F., Reis, R. M., and Simao, A. (2021). A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. *Scientific reports*, 11(1):1–10.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2019a). GeNWeMME: A network-based computational method for prioritizing groups of significant related genes in cancer. In *Advances in Bioinformatics and Computational Biology*, pages 29–40. Springer.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2019b). A proposal of a graph-based computational method for ranking significant set of related genes in cancer. In *Anais Principais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 300–305, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas.2019.6266.

Cutigi, J. F., Evangelista, A. F., and Simao, A. (2020a). Approaches for the identification of driver mutations in cancer: A tutorial from a computational perspective. *Journal of Bioinformatics and Computational Biology*, 18(03):2050016. PMID: 32698724. DOI: 10.1142/S021972002050016X.

Cutigi, J. F., Evangelista, R. F., Ramos, R. H., Ferreira, C. d. O. L., Evangelista, A. F., de Carvalho, A. C., and Simao, A. (2020b). Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In *Brazilian Symposium on Bioinformatics*, pages 81–92. Springer.

Das, J. and Yu, H. (2012). Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6:92. DOI: 10.1186/1752-0509-6-92.

Deng, Y., Luo, S., Deng, C., Luo, T., Yin, W., Zhang, H., Zhang, Y., Zhang, X., Lan, Y., Ping, Y., *et al.* (2019). Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. *Briefings in Bioinformatics*, 20(1):254–266.

Dentro, S. C., Wedge, D. C., and Van Loo, P. (2017). Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor Perspectives in Medicine*. DOI: 10.1101/cshperspect.a026625.

Dujon, A., Aktipis, C., Alix-Panabières, C., Amend, S., Boddy, A., Brown, J., Capp, J., DeGregori, J., Ewald, P., Gatenby, R., Gerlinger, M., Giraudeau, M., Hamede, R., Hansen, E., Kareva, I., Maley, C., Marusyk, A., McGranahan, N., Metzger, M., and Ujvari, B. (2021). Identifying key questions in the ecology and evolution of cancer. *Evolutionary Applications*, 14:877–892. DOI: 10.1111/eva.13190.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. DOI: 10.1109/SFCS.2000.892133.

Greaves, M. and Carlo, M. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313. DOI: 10.1038/nature10762.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., *et al.* (2020). The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic acids research*, 37(Database issue):D767–72. DOI: 10.1093/nar/gkn892.

Kim, Y., Cho, D., and Przytycka, T. M. (2016). Understanding genotype-phenotype effects in cancer via network approaches. *PLoS Computational Biology*, 12(3). DOI: 10.1371/journal.pcbi.1004747.

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., *et al.* (2020). A reference map of the human binary protein interactome. *Nature*, pages 1–7.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P. S., Papaemmanuil, E., Davies, H. R., Varela, I., McBride, D. J., Bignell, G. R., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerød, A., Aparicio, S. A., Tutt, A., Sieuwerts, A. M., Borg, A., Thomas, G., Salomon, A. V., Richardson, A. L., Børresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., and Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell (Cambridge)*, 149(5):994–1007.

Ozturk, K., Dow, M., Carlin, D. E., Bejar, R., and Carter, H. (2018). The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430(18):2875–2899.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10):2363–71.

Rabadan, R. and Blumberg, A. J. (2019). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press.

Ramos, R. H., Cutigi, J. F., de Oliveira Lage Ferreira, C., Evangelista, A. F., and Simao, A. (2020). Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In *Anais Principais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 37–48, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbcas.2020.11500.

Ramos, R. H., Cutigi, J. F., Oliveira Lage Ferreira, C. d., and Simao, A. (2021). Topological characterization of cancer driver genes using reactome super pathways networks. In *Brazilian Symposium on Bioinformatics*, pages 26–37. Springer.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127):1546–1558. DOI: 10.1126/science.1235122.