Journal of Information and Data Management, 2024, 15:1, doi: 10.5753/jidm.2024.2625

Bioinformatics of infectious and chronic diseases at the Center for Technological Development in Health of Fiocruz

Nicolas Carels [**Fundação Oswaldo Cruz** | ***nicolas.carels@fiocruz.br***]
Gilberto Ferreira da Silva [**Fundação Oswaldo Cruz** | ***gilberto.silva@fiocruz.br***]
Carlyle Ribeiro Lima [**Fundação Oswaldo Cruz** | ***carlyle.lima@fiocruz.br***]
Franklin Souza da Silva [**Fundação Oswaldo Cruz** | ***franklin.souza@fiocruz.br***]
Milena Magalhães [**Fundação Oswaldo Cruz** | ***milena.magalhaes@fiocruz.br***]
Ana Emília Goulart Lemos [**Fundação Oswaldo Cruz** | **INCA** | **PICTIS** | ***ana.goulart@fiocruz.br***]

Platform of Biological System Modeling, Center for Technological Development in Health, Fundação Oswaldo Cruz (Fiocruz), Av. Brasil 4365, Manguinhos, Rio de Janeiro, RJ, 21040-900, Brazil.
Program of Immunology and Tumor Biology, Brazilian National Cancer Institute (INCA), Rio de Janeiro, RJ, 20231-050, Brazil.
Plataforma Internacional para Ciência, Tecnologia e Inovação em Saúde (PICTIS), Edificio Central, Via do Conhecimento 3830-352, Ílhavo, Portugal.

Received: 4 April 2022 • **Published:** 17 February 2024

Abstract One of the bioinformatics purposes is data mining and integration to solve fundamental scientific challenges. We have been investigating biological systems including viruses, bacteria, fungi, protozoans, plants, insects, and animals with such concern. Gradually, we moved from basic questions on genome organization to application in infectious and chronic diseases by integrating interactome and RNA-seq data to modeling techniques such as Flux Balance Analysis, structural modeling, Boolean modeling, system dynamics, and computation biology in a system biology perspective. At the moment, we focus on the rational therapy of cancer assisted by RNA sequencing, network modeling, and structural modeling.

Keywords: Information retrieval, Data Mining and Integration, System Modeling

1 Introduction

Computational techniques and mathematical modeling have been increasingly utilized to address complex biological inquiries arising from raw data generated by sequencing machines, mass spectrometers, and other high-throughput technologies [D'Argenio, 2018]. These datasets, processed by bioinformatic pipelines, necessitated the creation of virtual entities to represent the intricacies of biological structures and processes. Consequently, programming languages and databases were adapted to support object-oriented applications, which are now accessible via the internet [NRCC-FICB, 2005; Kindler and Krivy, 2011; Grochowski *et al*., 2019]. The continuous advancement in electronic device performance has facilitated the emergence of new high-level programming languages characterized by higher levels of abstraction. These languages are more user-friendly and can seamlessly integrate with various libraries, enabling them to execute intricate operations[1]. Consequently, bioinformatics has established itself as an interdisciplinary domain dedicated to the development of methods and software tools tailored for comprehending extensive and intricate biological datasets. This interdisciplinary field amalgamates principles from biology, chemistry, physics, computer science, information engineering, mathematics, and statistics to meticulously analyze and interpret these datasets. It is within this framework that the transition from traditional bench experimentation to a new experimental paradigm involving bioinformatics investigation and inference has taken place.

The primary objective of the Laboratory of Biological System Modeling is to serve as a bioinformatics platform, offering services to the partners affiliated with the *Center for Technological Development in Health* (CDTS[2]) at Fiocruz[3]. CDTS was established in 2002, and although its main facility is currently under construction[4], research and development teams were established in 2009. CDTS is dedicated to translational science, focusing on the development of health-related products and processes. The competencies entrusted to CDTS by Fiocruz encompass: (i) Facilitating coordination, management, and promotion of technological development and innovation; (ii) Enhancing technological development and innovation in health-related products and processes; (iii) Providing services based on its platforms and supporting laboratories within a flexible structure (product incubator); and (iv) Prospecting, analyzing, and communicating strategic investigations in technological development and innovation in health. Initially, CDTS commenced its operations with 11 technological platforms, which are detailed on its portal[5]. In Rio de Janeiro, CDTS serves as a prototype institute, bridging the gap between the basic science conducted by the *Oswaldo Cruz Institute* (IOC) and Fiocruz' production units: Biomanguinhos (biopharmaceuticals) and

[1] https://en.wikipedia.org/wiki/History_of_programming_languages

[2] https://www.cdts.fiocruz.br/
[3] https://www.fiocruz.br/
[4] https://paal.com.br/blog/portfolio/cdts-2/
[5] https://www.cdts.fiocruz.br/linhas-de-pdi

Farmanguinhos (drugs). Notably, Fiocruz has recently undergone significant geographic expansion, and CDTS is expected to collaborate with Fiocruz's regional centers as part of its extended responsibilities.

2 History of the group and its members

Our Research Group[6] is led by Nicolas Carels who started its activities in Genomics in the early 1990s in the *Jacques Monod Institute* (IJM[7]) under the *Centre National de la Recherche Scientifique* (CNRS[8]). The primary objective was to investigate the physical distribution of genes in animal and plant systems. Initially, these investigations were conducted through bench experiments. During that period, GenBank, established in 1979, had descriptions for only a limited number of genes. Despite the limited data available, integration of the existing information was possible. Combined with bench results, this integration enabled us to address various scientific challenges. This fusion of digital and bench data marked the initiation of the bioinformatics era [Gauthier *et al*., 2019]. The field of bioinformatics advanced in tandem with high-throughput techniques for cloning, expressing, and characterizing biological molecules [D'Argenio, 2018].With the continual enhancement of electronic devices' performance, bioinformatics attained higher levels of complexity, facilitating the integration of increasingly sophisticated mathematical tools [Motta and Pappalardo, 2013]. Since then, our focus has been on data mining and integration to tackle fundamental scientific problems.

2.1 Genome structure and organization

2.1.1 Genome organization and nucleotide bias

Initially, our research delved into gene distribution and nucleotide biases in eukaryote genomes [Carels *et al*., 1998]. A noteworthy discovery emerged when examining the regional composition in guanine plus cytosine (*GC*). Surprisingly, the genome heterogeneity of intergenic sequences in humans exhibited greater diversity (5 compositional domains, [Costantini *et al*., 2006]) compared to maize (2 compositional domains, [Carels, 2005b]), despite both species having similar genome sizes (3 Gbp) and nucleotide biases in the third position of codons (30%-100%). Compositional domains (coined as isochores by Bernardi), defined as large genomic regions (>1 Mbp) homogeneous in *GC* in the 4% interval, illustrated that selective constraints applied differently to the intergenic sequences of plants and warm-blooded vertebrates [Carels, 2005a]. This observation indicated a distinctive genome phenotype resulting from different plants' genomic strategies in response to environmental pressures. Additionally, a compositional transition in coding sequences of plants was noted, particularly between dicotyledons and Poaceae [Carels *et al*., 1998]. Considering distinct genomic strategies between plants and animals, a similar intron size and number bias was observed between *GC*-poor (numerous long introns) and *GC*-rich (few short introns) genes in both plants and vertebrates, with the contrast being more pronounced in plants than in vertebrates [Carels and Bernardi, 2000]. The disparity in intron composition between *GC*-poor and *GC*-rich genes was found to be linked with *CpG* islands in gene promoter regions. *GC*-poor genes lacked *CpG* islands, while *GC*-rich genes were with an observed over expected frequency close to 1, indicating active maintenance (positive selection) despite the mutational bias of ^5mC into T [Carels, 2005a]. This finding led to the conclusion that *GC*-poor and *GC*-rich genes were subject to different types of regulation. Additionally, *GC*-rich genes tended to be expressed at higher levels than *GC*-poor genes, implying a network of functional correlation between genome organization and regional composition. These functional correlations were further demonstrated on a larger scale in the vertebrate model [Bernardi, 2021]. This research conducted at IJM and *Stazione Zoologica Anton Dohrn*[9] (Naples, Italy), prompted a closer examination of nucleotide composition in *coding sequences* (CDS). The study was initiated at the *Universidade Estadual de Santa-Cruz*[10] (UESC, Bahia) and concluded at Fiocruz (Rio de Janeiro).

2.1.2 Coding sequence validation

The *universal correlation* recognized a similar bias in the 2^{nd} and 3^{rd} position of codons in prokaryotes and eukaryotes [D'Onofrio *et al*., 1999]. Naturally, the guanine plus cytosine level in the 3^{rd} position of codons (GC_3) increases much more rapidly than that of GC_2 due to the wobble nature of the 3^{rd} position of codons. However, the analysis of nucleotide composition revealed a periodic distribution of nucleotides along CDSs following what has been termed the ancestral codon, denoted by the Rrr pattern (*R* is for a large purine frequency in 1^{st} position of codons and *r* is for a low purine frequency in the 2 other positions of codons) [Carels *et al*., 2009; Carels and Frias, 2013a]. This structural pattern is nontrivial as nucleotide distribution typically occurs randomly within the codon table, raising questions about the existence of such periodicity. This periodicity is induced by a network of selective constraints on protein functionality that is universal across prokaryotes and eukaryotes [Ponce de Leon *et al*., 2014]. However, the coexistence of the universal Rrr ancestral codon with the universal correlation, based on the *S*/*W* ratio (*S* for strong or *GC* and *W* for weak or *AT*), implied a complex network of correlation between nucleotide composition and their positions in codons, a phenomenon not present in non-coding sequences such as introns [Carels and Frias, 2013b]. This led to the question: What constraints could be shaping the structure of codons?

Upon comparing codons and amino acid frequencies, we found that these constraints resulted from: (i) the energy cost associated with amino acid synthesis, (ii) the spatial distribution of turns and helices in contact with the solvent, and (iii) the spatial distribution of β-sheets at the protein center [Ponce de Leon *et al*., 2014]. The ancestral codon Rrr can also be represented as RNY or RWY, and even Ggg. The

[6] http://dgp.cnpq.br/dgp/espelhogrupo/184801
[7] https://www.ijm.fr/en/3/home-page.htm
[8] https://www.cnrs.fr/en
[9] https://www.szn.it/index.php/en/
[10] http://www.uesc.br/

reason R_1 (frequency of A or G purines in 1^{st} position of codons) is larger than R_2 or R_3 is that the 1^{st} position of codons encodes for the energy cost of amino acid synthesis; indeed, glycine, alanine, and valine are the amino acids with the simplest moieties and are also the most frequent in turns, helices and β-sheets, respectively. W_2 represents weak, i.e., A_2 or T_2, because T_2 is linearly correlated to hydropathy. Large T_2 values correspond to amino acids with hydrophobic moieties, which are typically much more expensive for a cell to synthesize than the hydrophilic ones found more commonly in turns and helices [Ponce de Leon *et al*., 2014]. The ancestral codon is characterized by a scarcity of purines in the second position of codons (r_2) with A_2 compensing G_2, while C_2 compensate for T_2, allowing T_2 to be correlated with hydropathy. However, GC_2 (S_2) is smaller than AT_2 (W_2) regardless of the organism under consideration being *GC*-poor or *GC*-rich. According to the RWY or Ggg pattern of the ancestral codon, pyrimidines (Y) are more frequent in the 3^{rd} position of codons. However, due to constraints on r_3 and T_2, C_3 compensates for base composition according to the *GC*-rich bias (or T_3 for the *GC*-poor bias, [Carels and Frias, 2013b]). This codon structure is recognized by Hidden Markov Model (HMM) models, but the Viterbi algorithm typically has a 15% error rate in reading frame identification because it does not includes any restriction for possible confusion of nucleotide probabilities between +1 and -1 (*GC* or *CG*: the condition $G_1 > G_2$ was true for all proteins cristallized and whose models were deposited in PDB[11] in 2013) or even between +1 and -2 (the condition A_1>T_2 was true for all proteins from PDB) positions [Carels and Frias, 2013a]. Implementing these restrictions increased the prediction rate of the coding frame among the six frames from *expressed sequence tags* (EST) contigs above 95% in most organisms, provide that the sequence size was at least 300 bp, which is true for 95% of GenBank proteins [Carels and Frias, 2013a]. These rules were combined in a statistical measure (UFM) that is independent of the nucleotide composition of a sequence and does not require any previous training step [Carels and Frias, 2013a]. This justifies why starting by annotating the exome is the first step to be performed when describing a new genome. The term "ancestral codon" originates from the facts that (i) a primeval codon system encoding Gly, Ala, and Val was sufficient to produce catalytic proteins with turns, helices, and β-sheets, provided that metallic ions would be chelated, and (ii) these amino acids were available in the primeval Earth conditions, as proven by Miller and successors (see in [Carels and Ponce de Leon, 2015]).

2.2 Plant genetics

The investigation into genome characterization led us to explore economically significant plant systems, such as *Theobroma cacau* (cacao for chocolate production), *Jatropha curcas* (physic nut for biodiesel production) as well as plant tolerance to drought. Plant research also held importance for medicinal purposes, an area of ongoing research at Farmanguinhos (Fiocruz). Initially, due to historical reasons, we commenced our bioinformatics research in plant systems in France and continued this work during our activities at UESC and Fiocruz. This approach served as a means of comparison with animal eukaryotes.

At UESC, our research focused on cacao, where we investigated *simple sequence repeats* (SSR) markers derived from ESTs of tissues infected by *Moniliophthora perniciosa*, a fungus causing witches' broom disease (WBD) in cacao. We compared the polymorphism of these EST-SSR with classical neutral SSR markers. Our study revealed several of these EST-SSR markers that are applicable for selective breeding of cacao [Lima *et al*., 2010].

In our research on physic nut, initiated at UESC, we conducted sequencing and characterization of a cDNA library from *J. curcas* L seeds at three stages of fruit maturation before yellowing occurred. The obtained sequences exhibited minimal redundancy, providing extensive metabolic coverage when compared through homology analysis with Gene Onthology (GO, [The Gene Ontology Consortium, 2008]). By comparing these ESTs and those available from GenBank with the *Kyoto Encyclopedia of Genes and Genomes* (KEGG[12]) [Kanehisa and Goto, 2000], we identified tags with nucleotide variations among *J. curcas* accessions for genes involved in fatty acid, terpene, alkaloid, quinone, and hormone biosynthetic pathways. Specifically, we assessed the expression levels of four genes (encoding palmitoylacyl carrier protein thioesterase, 3-ketoacyl-CoA thiolase B, lysophosphatidic acid acyltransferase, and geranyl pyrophosphate synthase) using real-time PCR. These genes exhibited significant differences in expression between leaves and fruits. Given that nucleotide polymorphisms in these genes are associated with higher expression levels in fruits compared to leaves, we proposed this approach for generating genetic markers to expedite the identification of *quantitative traits* (QTL) in selective breeding of *J. curcas* [Gomes *et al*., 2010].

Our research on plant drought tolerance is still in its early stages and mainly consists of a literature review. This review prompted the proposal of apomixis to stabilize epigenetic traits in wheat, an important feeding resource worldwide [Adel and Carels, 2023]. However, this topic is gaining global significance due to climate change, particularly in vulnerable regions.

2.3 Host pathogene interactions

2.3.1 The system of *Moniliophthora perniciosa* vs cacao

M. perniciosa is a fungal parasite of cacao that belongs to Basidiomycota. Studying the molecular relationship between the fungal parasite *M. perniciosa* and its host, *Theobroma cacau*, posed significant challenges. *M. perniciosa* infects the intercellular spaces of cacao meristematic tissues, with a ratio of *M. perniciosa* to *T. cacau* not exceeding 5%. This ratio approximates the error rate of a 5'-3' directional cloning kit, implying that under optimal conditions, not all ESTs could be expected to be cloned in the 5'-3' orientation. Moreover, these kits were costly, and most EST libraries did not consider the 5'-3' orientation [Gesteira *et al*., 2007]. Consequently, any read sequence might be coding or not, on the

[11] https://www.rcsb.org/

[12] http://www.genome.jp/kegg/

leading or lagging strand, indicating that if they were coding, they were in one of the six frames. It was not feasible to employ the same algorithm to determine whether a read was coding, identify its coding frame, and ascertain whether it originated from cacao or *M. perniciosa*, as genomic sequences for either organism were unavailable at that time. In addition, the use of a subtractive library did not seem to solve the problem [Garcia *et al*., 2011].

The resolution adopted involved classifying reads as either coding or non-coding and identifying their coding frame, if applicable, using the universal ancestral codon as a basis [Carels and Frias, 2013a]. Subsequently, a classifier was trained utilizing a ninth-order Markov model. The observed frequency was then compared with the expected frequencies derived from two matrices of 9-mers (one specific to cacao and the other to *M. perniciosa*) as references in the training sets. This system had an approximate error rate of 10% (unpublished data).

2.3.2 The system of *Fusarium oxysporum* vs plants

Continuing the exploration of potential molecular targets in fungal plant diseases, our focus shifted to the genus *Fusarium*, which comprises some of the most extensively studied and economically impactful plant pathogenic species in global agriculture and horticulture [Burgess and Bryden, 2012]. *Fusarium* spp. are widespread fungi [Agrios, 2004], found in soil, plants, various organic substrates, and are also recognized as opportunistic human pathogens [Zhang *et al*., 2020]. We hypothesized that pinpointing specific enzymes vital to *Fusarium* spp.'s metabolism might yield potential molecular targets for controlling the diseases they inflict on their hosts. Employing conventional methods such as sequence homology comparison through similarity search and Markov modeling, we characterized enzymatic functionalities associated with protein targets, which could be potential candidates for controlling root rots caused by *Fusarium oxysporum*. By comparing *F. oxysporum*'s specific enzymes with the genomes of *Arabidopsis thaliana*, *Brassica rapa*, *Glycine max*, *J. curcas*, and *Ricinus communis*, we identified a limited number of key enzymes. Inhibiting these enzymes was expected to significantly impact the fungus's development [Catharina and Carels, 2018]. The application of Flux Balance Analysis (FBA) modeling enabled the identification of a set of critical *F. oxysporum* enzymes essential for biomass production, implying that inhibiting them could potentially disrupt the fungus's metabolic network *in vivo*. Among these pivotal genes, F9F4G5 and F9FSB6 were identified as counterparts analogous to those in *A. thaliana*. Notably, only one pair's 3D structure could be modeled (F9F4G5 vs. Q8GWP5). Consequently, the enzyme F9F4G5 emerges as a promising target for inhibiting *F. oxysporum* since it plays a role in fungal growth and the regulation of key biological processes [Catharina, 2017][13].

2.3.3 The system of *Leishmania major* vs humans

Subsequently, our focus shifted to human parasites, specifically *Leishmania major*. Our objective was to identify enzymes unique to *Leishmania major* compared to *Homo sapiens*, with the potential to be considered targets for facilitating drug development. This approach relied on conventional techniques such as sequence homology comparison using similarity search (BLAST) and Markov modeling. Markov modeling provided the advantage of integrating the characterization of enzymatic functionality, secondary and tertiary protein structures (3D), protein domain architecture, and metabolic context. Through the identification of 42 enzymatic activities specific to *L. major* in comparison to *H. sapiens*, as classified by the *Analogous Enzymes Pipeline* (AnEnPi) [Otto *et al*., 2008], we pinpointed sterol 24-C-methyltransferase, pyruvate phosphate dikinase, trypanothione synthetase, and RNA-editing ligase as four essential enzymes for *L. major*. These enzymes could potentially serve as targets for drug development efforts [Catharina *et al*., 2017].

2.3.4 *In silico* structural characterization of protein targets from *Trypanosoma cruzi*

In our exploration of Trypanosomatidae family parasites, we turned our attention to *T. cruzi*, which is a significant human protozoan parasite in Brazil. Despite numerous experimental studies, effective treatments for Chagas disease remained elusive. Therefore, we conducted *in silico* predictions of the 3D structures of *T. cruzi* sequences that were either analogous to human proteins or unique to *T. cruzi*. These sequences were potential candidates for drug development. The identification of these protein targets was accomplished using the AnEnPi pipeline in a prior investigation [Gomes *et al*., 2011]. Analogous enzymes in *T. cruzi* were associated with trypanothione reductase, cysteine synthase, and ATPase, while sequences specific to *T. cruzi*, i.e., absent in *H. sapiens*, were associated with 2,4-dienoyl-CoA reductase and leishmanolysin activities. Our refined models, scrutinized through atomistic molecular dynamics (monomer or dimer) simulations in aqueous or bi-membrane solutions *in silico*, indicated that all protein targets, except cysteine synthase, warranted further investigation [Lima *et al*., 2016].

2.3.5 The system of *T. cruzi* in triatomines

At that time, extensive research had been conducted on *T. cruzi* in humans, but very limited information was available regarding its interaction with bacteria in the *digestive tract of triatomines* (DTT), its insect vectors. As *T. cruzi* was known to reside within the DTT without breaching its cell boundary, we posed a question: could the composition of the triatomine bacterial microbiota influence its population? Given the insufficient description of this system and the scant available data, which indicated the prevalence of *Serratia marcescens* in *in vitro* culture, we opted to explore its microbial composition by sequencing 16S rDNA. This choice was motivated by the potential presence of bacterial species that might be unculturable. We deemed it crucial to investigate the triatomine microbiota because the triatomine gut serves as one of the factors that could impact the transmission and virulence of *T. cruzi* in humans. Understanding the microbiota composition, particularly in different triatomine species, could prove

[13] https://www.arca.fiocruz.br/handle/icict/37782

instrumental in devising strategies for biological control of *T. cruzi*.

1. Microbiota of digestive tract in Triatomines:
   In our analysis of the predominant bacterial species within the digestive tract (DTT) of *Rhodnius*, *Triatoma*, *Panstrongylus*, and *Dipetalogaster* genera using 16S rDNA sequencing, we observed a limited diversity of bacterial species, with fewer than 20 predominant species, as indicated by the Chao index. The composition of these species varied among different triatomine species. Specifically, our findings revealed that (i) *Serratia* was prevalent in *Rhodnius*, (ii) *Arsenophonus* was prevalent in *Triatoma* and *Panstrongylus*, while (iii) *Candidatus Rohrkolberia* was the primary species in *Dipetalogaster*. Unlike certain insect systems such as termites, the bacterial microbiota in triatomines exhibited a low complexity, with its structure differing based on the vector genus. Similar conclusions were drawn by other researchers studying hematophagous insects like mosquitoes [da Mota *et al*., 2012]. In a subsequent study, we examined triatomines captured in Ceará. The bacterial community in the gut of these peridomestic triatomines, identified as *Triatoma pseudomaculata* and *Triatoma brasiliensis* through COI sequence comparison, was predominantly composed of Proteobacteria and Actinobacteria. Additionally, Firmicutes and Bacteroidetes were present, although in lower proportions. The predominant genus remained *Serratia*, with members of Corynebacterinae, a suborder of Actinomycetales, forming the next significant group [Gumiel *et al*., 2015].

2. Metagenomics:
   Based on the preceding reports, the following hypothesis-consequence relationships were formulated: (i) Considering the DTT as an ecological niche supporting microbiota adapted to specific substrate availability, we investigated the molecular enzymatic properties favoring bacterial prominence in this environment; (ii) Utilizing the microbiota composition of DTT from previous 16S rDNA analyses and whole sequenced genomes of bacteria within the same genera available in GenBank, we calculated the *GC* content of rare and prominent bacterial species in the DTTs; and (iii) Recognizing the conservation of genome *GC* content within bacterial genera [Takahashi *et al*., 2009], we compared the enzymatic reactions encoded by CDSs of both rare and common bacterial species, elucidating key functions explaining the competitive advantage of certain genera in the DTT.
   Therefore, we examined the composition of DTT microbiota by conducting shotgun sequencing of DNA extracted from bacteria cultured in liquid Luria-Bertani broth (LB) medium. The findings revealed that bacteria with a high *GC* content effectively outperformed those with a low *GC* content, establishing their dominance within the DTT microbiota.
   The comparison of genes sequence from the bacteria previously listed with KEGG showed that oxidoreductases were the main enzymatic components of DTT microbiota. Specifically, nitrate reductases (involved in anaerobic respiration), oxygenases (for the catabolism of complex substrates), acetate-CoA ligase (related to the tricarboxylic acid cycle and energy metabolism), and kinase (associated with signaling pathways) were the predominant enzymatic determinants. These enzymes were accompanied by a variety of minor enzymes, including hydrogenases involved in energy and amino acid metabolism. We concluded that bacteria from *GC*-rich genera outcompete those of *GC*-poor ones due to their specific enzymatic capabilities, providing them with a selective advantage in the DTT for the catabolism of complex molecules such as hemoglobin [Carels *et al*., 2017].

3. Metabolomics:
   As mentioned earlier, the composition of microbiota differs not only between triatomine species but also across different blood meals. Consequently, the dynamic nature of the DTT, where *T. cruzi* resides, can significantly impact its development. Recognizing this, we embarked on a study to explore the chemical composition of the DTT using a metabolomics approach. Utilizing *Direct Infusion Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*, we analyzed fecal samples from three triatomine species (*Rhodnius prolixus*, *Triatoma infestans*, *Panstrongylus megistus*) following rabbit blood meals. Subsequently, we identified clusters of metabolites that were either consistently present in all species (contingent core) or specifically enriched in each species (specific core). By querying the *Human Metabolome Database*[14] [Wishart *et al*., 2007], we determined putative identities for these metabolites of interest. Our findings revealed that approximately 80% of the detected molecules constituted a fundamental set of metabolites present uniformly across all species, while the remaining 20% exhibited variations among the triatomine species. The contingent core set of metabolites encompassed diverse categories such as fatty acids, steroids, glycerolipids, nucleotides, and sugars, among others. Nevertheless, the metabolic signature of triatomine feces also exhibited variations among the specific species under consideration. The contingent core primarily comprised prenol lipids, amino acids, glycerolipids, steroids, phenols, fatty acids and derivatives, benzoic acid and derivatives, flavonoids, glycerophospholipids, benzopyrans, and quinolines. Our findings lead us to infer that the abundant and diverse chemical components within the DTT milieu are likely to influence the development and infectivity of *T. cruzi*. The intricate nature of the fecal metabolome in triatomines implies that it could impact the vector competence of triatomines for specific *T. cruzi* strains. Understanding the chemical environment of *T. cruzi* within its invertebrate host holds the potential to reveal novel insights into the factors influencing parasite proliferation and provide

[14] https://hmdb.ca/

strategies for Chagas disease control [Antunes *et al*., 2013].

4. Proteomics:
   As mentioned earlier, *R. prolixus*, *P. megistus*, *T. infestans*, and *D. maxima* are all triatomines and potential vectors of *T. cruzi*, the causative agent of human Chagas' disease. Recognizing that the life cycle of *T. cruzi* unfolds within the DTT, we deemed it imperative to analyze the protein profile of the DTT as a crucial step in comprehending the physiology of the DTT during *T. cruzi* infection. To delineate the protein profile of DTT in *D. maxima*, *P. megistus*, *R. prolixus*, and *T. infestans*, we conducted a comprehensive analysis using shotgun liquid chromatography-tandem mass spectrometry (LC-MS/MS). The majority of identified proteins were closely associated with metabolic pathways such as gluconeogenesis/glycolysis, citrate cycle, fatty acid metabolism, oxidative phosphorylation, and immune responses. This novel proteomic dataset was annotated and integrated with previously published data in alignment with Gene Ontology (GO) and KEGG classifications. Enzymes were categorized based on class, acceptor, and function, while proteins related to the immune system were annotated with reference to pathways including humoral response, cell cycle regulation, Toll, IMD, JNK, Jak-STAT, and MAPK, as available from the *Insect Innate Immunity Database* (IIID). The identified pathways were further categorized into recognition, signaling, response, coagulation, melanization, and non-specific categories. Additionally, phylogenetic relationships and gene expression patterns of annexins were examined to comprehend their role in safeguarding and maintaining the intestinal epithelial cells against inflammation [Gumiel *et al*., 2020]. Through these comprehensive approaches encompassing metagenomics, metabolomics, and proteomics, we have significantly contributed to illuminating the environment in which *T. cruzi* thrives in its invertebrate hosts. This study has given rise to numerous questions, inspiring our colleagues to delve deeper into the relationship between *T. cruzi* and DTT. Consequently, this enhanced understanding represents a significant stride toward advancing the control measures for Chagas' disease.

3 Present times

3.1 Viral diseases

The investigation of viral diseases commenced with the analysis of mutation rates among dengue strains. It was observed that dengue 2 exhibited a higher mutation rate, determined by aligning numerous sequences of the four types [de Araújo *et al*., 2007]. During the COVID-19 pandemic, in collaboration with the research teams of Thiago Moreno L. Souza and Salvatore Giovanni De-Simone (Fiocruz/CDTS), our focus returned to viral diseases, this time concentrating on *in silico* protein modeling utilizing structural biology techniques (biophysics). Through *in silico*, structural analyses, we characterized the interaction between heme-binding motifs and hemin in several proteins, including the nucleoprotein (N), spike (S), core membrane protein (M), and non-structural proteins (Nsp3 and Nsp7) of SARS-CoV-2 [Lechuga *et al*., 2021].

We also conducted an investigation on the major protease (Mpro) of SARS-CoV-2, recognized as a promising drug target. We compared the *in silico* compatibility of atazanavir, lopinavir, and ritonavir (three antiretrovirals) with this protein. These analyses provided valuable data to supplement *in vitro* studies, leading to the suggestion that atazanavir and the combination of atazanavir with ritonavir should be regarded as potential candidate drugs for repurposing in the ongoing clinical trials against COVID-19 [Fintelman-Rodrigues *et al*., 2020].

3.2 Cancer diseases

With the improvement of economic conditions in Brazil, the health issues prevalent among its population are shifting from dysregulations caused by parasites to dysregulations induced by physiological changes, notably cancer. The increasing incidence of cancer is anticipated to significantly impact the healthcare system, particularly due to the costly palliative cares. As a result, our research focus has transitioned from parasitic diseases to cancer-related studies.

3.2.1 Hub targeting in malignant network

Cancer is a complex disease resulting from genetic and epigenetic changes, which disrupts several mechanisms, including cell division regulation. We hypothesized that targeting up-regulated protein hubs within the framework of theranostics, combining the diagnosis and therapeutic approach to cancer, could enhance patients' benefits. The advantage of targeting hubs in scale free networks has been mathematically demonstrated by the Barabasi's research team [Albert *et al*., 2000; Barabási, 2016][15] and confirmed by Conforte *et al*. [2019], encompassing various types of cancer with different degrees of aggressiveness (ranging from 30% to 98% 5-years overall survival rates). Through an analysis of the entropy [Shannon, 1948] of signaling subnetwork of up-regulated genes in malignant cells, conducted by comparing the IntAct[16] interactome [Orchard *et al*., 2014] and RNA-seq data from *The Cancer Genome Atlas* (TCGA[17]), we identified hubs that needed inhibition. Notably, these hubs displayed varying levels of specificity concerning different tumor types and individual patients, underscoring the necessity for a personalized approach to rational therapy rather than a generic *one-size-fits-all* solution [Carels *et al*., 2015]. This concept was further substantiated through *in vitro* experiments on breast cancer cell lines using RNA interference, as demonstrated by Tilli *et al*. [2016] and was later extended to tumors by Conforte *et al*. [2019]. The research conducted by Conforte *et al*. [2019] expanded the rational therapy paradigm, focusing on inhibiting up-regulated hubs, from cellular models to tumor

[15] http://barabasi.com/networksciencebook/
[16] https://www.ebi.ac.uk/intact/
[17] https://cancergenome.nih.gov/

scenarios through RNA-seq data analyses. Their study revealed a correlation between the number of hubs to be deactivated and tumor aggressiveness. Specifically, there was a negative correlation between entropy and 5-year overall survival rates. This correlation enabled the calculation that approximately 10 hubs should be inhibited for tumors with a 30% 5-years overall survival rate, while about 3 hubs should be targeted for tumors with a 98% 5-years overall survival rate, on average. Intermediate values can be easily derived due to the linear regression that can be fitted on the observed data of the negative correlation as indicated by Conforte *et al*. [2019]. The negative correlation has also enabled to derive genes responsible for tumor aggressiveness as a result of *principle component analysis* (PCA) and *random forest classification* [Barbosa-Silva *et al*., 2022].

The hub approach is oriented toward the molecular phenotype, whereas the traditional method is based on gene mutations and statistical analysis of large patient cohorts. The distinction between these two approaches lies in the direct connection of the patient's disease condition, under the hub-based strategy, to a phenotypic portrait of molecular dysfunction, enabling a rational therapy. In contrast, the mutation-based approach depends on an indirect relationship between marker genes and the disease itself through statistical correlations. This dependency assumes a significant variability in the disease response to drug treatments.

This research resulted in the approval of a patent (BR102015030819-1) by the *Brazilian Institute of Intellectual Property Protection* (INPI[18]). It has been financially supported by Faperj[19] to facilitate its transition to patients in a start-up context. Consequently, we automated the hub diagnosis, which was previously conducted using a series of Perl and Python scripts, within a Galaxy pipeline accessible through web pages [Pires *et al*., 2021]. Currently, efforts are underway to advance toward *in vivo* pre-clinical validation in the context of solid and liquid tumors in collaboration with Flávia Raquel Gonçalves Carneiro (Fiocruz/CDTS) and Ana Carolina dos Santos Monteiro from *Laboratório de Osteoimunologia e Imunologia Tumoral* (Universidade Federal Fluminense - UFF). We are also looking for biomarker of HPV-induced cervix cancer staging in collaboration with Cecilia Vianna de Andrade from *Coordenação Diagnóstica em Anatomia Patológica e Citopatologia* (Fiocruz/Instituto Fernandes Figueira - IFF).

3.2.2 Cancer modeling

Despite the promising outlook presented by our hub-based approach to cancer treatment, we recognized the presence of challenging inquiries that needed exploration. One such query involved determining whether it would be more effective to target hubs that are neighbors in the up-regulated network of tumors, or hubs that are distant in such networks. This led us to explore a gene targeting method based on the concept of *attractors*. These questions are currently under investigation in collaboration with the research team led by Fabricio Alves Barbosa da Silva at the *Laboratory of Computational Modeling of Biological Systems* (Fiocruz/Programa de Computação Científica - PROCC).

1. Hopfield networks:
   We initially addressed the issues outlined above using *Hopfield networks*. Considering the possibility that cancer may correspond to attractors in *Waddington's epigenetic landscape* [Huang *et al*., 2009], employing Hopfield networks to model basins of attraction presented an appealing approach. This method did not necessitate prior biological knowledge about protein-protein interactions or kinetic parameters. Thus, we utilized Hopfield network modeling to analyze bulk RNA-seq data from paired tumor and control samples from breast cancer. We characterized the attractors of the control and tumor samples in terms of their size and potential energy. Subsequently, we examined the correlation between the Euclidean distances among the tumor samples and the control attractor with their respective clinical data. Our findings revealed that the tumor basin of attraction was larger than the control one and that tumor samples exhibited significantly more negative energy than control samples, aligning with prior research findings [Conforte *et al*., 2020].

2. Boolean modeling:
   An alternative approach involved modeling the up-regulated malignant networks using Boolean techniques. However, this method had a drawback: it required knowledge of edge orientation and function (activation or inhibition), which might not always be available. Understanding the activation or inhibition status of the genes within the network during its temporal evolution is crucial for intervening rationally in controlling the system's dynamic changes. Consequently, we proposed a methodology for constructing data-driven Boolean networks that represent breast cancer tumors. In this approach, we defined the network components and topology based on gene expression data derived from RNA-seq of breast cancer cell lines. We employed a Boolean logic formalism to describe the dynamics of the network. The integration of single-cell RNA-seq (scRNA-seq) and interactome data provided us with the opportunity to investigate the dynamics of malignant subnetworks consisting of up-regulated genes through canalyzing functions. Utilizing single-cell breast cancer datasets sourced from TCGA, we employed a binarization algorithm. This transformed version of scRNA-seq data facilitated the identification of attractors specific to individual patients and critical genes associated with each subtype of breast cancer [Sgariglia *et al*., 2021]. This model has been investigated to detect critical genes involved in malignant attractor stability whose inhibition could optimize the induction of tumor cell death and serve for potential applications in cancer theranostics [Sgariglia *et al*., 2023]. The proposed model is currently under development and holds the potential to form the basis for a methodology that aims at identifying pivotal genes involved in the stability of malignant attractors. Inhibiting these genes could have significant implica-

[18] http://www.gov.br/inpi/pt-br
[19] http://www.faperj.br/

tions in the field of cancer theranostics.

3. Cancer system dynamics:
   Employing single-cell RNA sequencing data, we conducted a comprehensive analysis of the cellular landscape in *Glioblastoma Multiforme* (GBM). Drawing inspiration from observed traits in stochastic systems, we introduced factors such as genomic instability as a noise source, effectively characterizing cancer dynamics through stochastic fixed points. We treated sample and time averages as equivalent, aiding both in parameter fitting and subsequent stochastic simulations of cluster centroids. This methodology gained support through the correlation found between centroids of experimental and simulated datasets. The use of stochastic modeling to delineate the Waddington landscape improved our understanding of GBM's cellular heterogeneity and enabled a visual framework for validating the centroids as accurate representations of potential cancer attractors. Specifically, this approach bridged the gap between molecular-level variations and the basin of attraction. Additionally, we explored the stability and transitions between the attractors linked to different subtypes of GBM [Vieira *et al*., 2023].

3.3 Drug repurposing

The identification of protein hubs for rational therapy, coupled with *in silico* modeling of protein targets, naturally leads to the proposition of drug repurposing. Through the rational diagnosis of protein hubs in solid tumors across nine cancer types, we pinpointed ~100 potentially significant genes, with ~60 being particularly pertinent and ~30 having corresponding PDB models. Having protein models available for these potentially crucial cancer targets, along with drug structures in ZINC[20] [Irwin *et al*., 2012] and data on approved drugs in DrugBank[21] [Wishart *et al*., 2017], enables us to conduct screenings for drugs that align with the theranostics concept (Lima et al. unpublished data).

4 Conclusion

The use of integrated methods of bioinformatics, computational biology, and artificial intelligence in biological sciences is a rapidly growing field and is anticipated to provide benefits to the society, professionals, governments, and insurance companies. The journey of our research team as presented in this report aligns with such trend including in health sciences since rational chemotherapy has the potential to reduce patient suffering, enhance diagnostic accuracy, and reduce costs, all of which would be advantageous for society as a whole.

Acknowledgements

We would like to acknowledge the colleagues who the most actively participated to our journey: *Institut Jacque Monod* (IJM): G. Bernardi, O. Clay, E. Geigl, *Università degli Studi di Napoli Federico II*: G. Saccone, *Centro de Astrobiología* (CAB) from *Instituto Nacional de Técnica Aeroespacial* (INTA): Juan Pérez-Mercader, *Universidade Estadual de Santa Cruz* (UESC): D. Frias, F. Micheli, J. Cascardo, *University of Alberta* (UofA): J. Tuszynski, *Fiocruz*: A. B. de Miranda, F. A. B. da Silva, W. Degrave, T. Souza, S. De Simone, C. Andrade, C. Morel.

Funding

We were funded by Centre National de la Recherche Scientifique (CNRS, France), Consiglio Nazionale delle Ricerche (CNR, Italy), Consejo Superior de Investigaciones Científicas (CSIC, Spain), Fundação de Amparo à Pesquisa do Estado da Bahia (Fapesb, Brazil), Fundação de Amparo à Pesquisa do Estado de Rio de Janeiro (Faperj, Brazil), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil), and University of Alberta (UofA, Canada) along this journey.

Competing interests

The authors declare that they have no competing interests.

[20] https://zinc12.docking.org/

[21] https://go.drugbank.com/

References

Adel, A. and Carels, N. (2023). Plant tolerance to drought stress with emphasis on wheat. *Plants*, 12(11):2170.

Agrios, G. N. (2004). *Plant Pathology*. Elsevier.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.

Antunes, L. C. M., Han, J., Pan, J., Moreira, C. J. C., Azambuja, P., Borchers, C. H., and Carels, N. (2013). Metabolic signatures of triatomine Vectors of *Trypanosoma cruzi* unveiled by metabolomics. *Plos One*, 8:e77283–12.

Barabási, A.-L. (2016). *Network Science*. CambridgeUniversityPress.

Barbosa-Silva, A., Magalhães, M., da Silva, G. F., da Silva, F. A. B., Carneiro, F. R. G., and Carels, N. (2022). A data science approach for the identification of molecular signatures of aggressive cancers. *Cancers*, 14(9):2325.

Bernardi, G. (2021). The "Genomic Code": DNA pervasively moulds chromatin structures leaving no room for "Junk". *Life*, 11(4):342.

Burgess, L. W. and Bryden, W. L. (2012). *Fusarium*: a ubiquitous fungus of global significance. *Microbiol Australia*, 33(1):22.

Carels, N. (2005a). *Recent Research Developments in Plant Science*, chapter The genome organization of angiosperms, pages 129–194. Research Signpost.

Carels, N. (2005b). The maize gene space is compositionally compartimentalized. *FEBS Letters*, 579(18):3867–3871.

Carels, N. and Bernardi, G. (2000). Two classes of genes in plants. *Genetics*, 154:1819–1825.

Carels, N. and Frias, D. (2013a). A statistical method without training step for the classification of coding Frame in transcriptome sequences. *Bioinf Biol Insights*, 7:35–54.

Carels, N. and Frias, D. (2013b). *BIOMAT 2012*, chapter The contribution of stop codon frequency and purine bias to the classification of coding sequences, pages 301–322. World Scientific.

Carels, N., Gumiel, M., da Mota, F. F., Moreira, C. J., and Azambuja, P. (2017). A metagenomic analysis of bacterial microbiota in the digestive tract of triatomines. *Bioinf Biol Insights*, 11:117793221773342.

Carels, N., Hatey, P., Jabbari, K., and Bernardi, G. (1998). Compositional properties of homologous coding sequences from plants. *J Mol Evol*, 46:45–53.

Carels, N. and Ponce de Leon, M. (2015). An interpretation of the ancestral codon from Miller's amino acids and nucleotide correlations in modern coding sequences. *Bioinf Biol Insights*, 9:37–47.

Carels, N., Tilli, T., and Tuszynski, J. A. (2015). A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases. *Plos One*, 10:e0115054–16.

Carels, N., Vidal, R., and Frías, D. (2009). Universal features for the classification of coding and non-coding DNA sequences. *Bioinf Biol Insights*, 3:37–49.

Catharina, L. and Carels, N. (2018). Specific enzyme functionalities of *Fusarium oxysporum* compared to host plants. 676:219–226.

Catharina, L., Lima, C. R., Franca, A., Guimarães, A. C. R., Alves-Ferreira, M., Tuffery, P., Derreumaux, P., and Carels, N. (2017). A computational methodology to overcome the challenges associated with the search for specific enzyme targets to develop drugs against *Leishmania major*. *Bioinf Biol Insights*, 11:1177932217712471.

Catharina, L. C. (2017). *Busca de alvos proteicos para desenvolvimento de inibidores enzimticos em sistemas hospedeiros-parasitos*. PhD thesis, Fiocruz, Rio de Janeiro, Brazil.

Conforte, A. J., Alves, L., Coelho, F. C., Carels, N., and da Silva, F. A. B. (2020). Modeling basins of attraction for breast cancer using Hopfield networks. *Front Genetics*, 11:314–317.

Conforte, A. J., Tuszynski, J. A., da Silva, F. A. B., and Carels, N. (2019). Signaling complexity measured by Shannon entropy and its application in personalized medicine. *Front Genetics*, 10:1–14.

Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Res*, 16(4):536–541.

da Mota, F. F., Marinho, L. P., Moreira, C. J., Lima, M. M., Mello, C. B., Garcia, E. S., Carels, N., and Azambuja, P. (2012). Cultivation-independent methods reveal differences among bacterial gut microbiota in triatomine vectors of Chagas disease. *PLos Negl Trop Dis*, 6(5):e1631.

D'Argenio, V. (2018). The high-throughput analyses era: Are we ready for the data struggle? *High Throughput*, 7(1):8.

de Araújo, R. F. S., Carels, N., de Melo, P. R. S., and FRIAS, D. (2007). Investigation of polymorphisms in the genome of Dengue virus. *RECIIS*, 1:316–320.

D'Onofrio, G., Jabbari, K., Musto, H., and Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, 238(1):3–14.

Fintelman-Rodrigues, N., Sacramento, C. Q., Lima, R. C., Souza da Silva, F., Ferreira, A. C., Mattos, M., de Freitas, C. S., Cardoso Soares, V., da Silva, G. D. S., Temerozo, J. R., Miranda, M. D., Matos, A. R., Bozza, F. A., Carels, N., Alves, C. R., Siqueira, M. M., Bozza, P. T., and Souza, T. M. L. (2020). Atazanavir, alone or in combination with ritonavir, inhibits SARS-CoV-2 replication and proinflammatory cytokine production. *Antimicrob Agents Chemother*, 64:e00825–20.

Garcia, D., Carels, N., Koop, D. M., de Souza, L. A., Andrade, J. S., Pujade-Renaud, V., Mattos, C. R. R., and Cascardo, J. C. M. (2011). EST profiling of resistant and susceptible Hevea infected by *Microcyclus ulei*. *Physiol Mol Plant Pathol*, 76:126–136.

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A brief history of bioinformatics. *Brief Bioinform*, 20(6):1981–1996.

Gesteira, A. S., Micheli, F., Carels, N., Silva, A. C. D., Gramacho, K. P., SCHUSTER, I., Macedo, J. N., Pereira, G. A. G., and Cascardo, J. C. M. (2007). Comparative analysis of expressed genes from cacao meristems infected by *Moniliophthora perniciosa*. *Ann Bot*, 100:129–140.

Gomes, K. A., Almeida, T., Gesteira, A., Lobo, I. P., Guimarães, A. C. R., de Miranda, A. B., Sluys, M.-A. V., da Cruz, R. S., Cascardo, J. C. M., and Carels, N. (2010). ESTs from seeds to assist the selective breeding of *Jatropha curcas* L. for oil and active compounds. *Genom Insights*, 3:29–56.

Gomes, M. R., Guimarães, A. C. R., and de Miranda, A. B. (2011). Specific and nonhomologous isofunctional enzymes of the genetic information processing pathways as potential therapeutical targets for Tritryps. *Enzyme Res*, 2011:1–8.

Grochowski, K., Breiter, M., and Nowak, R. (2019). *Introduction to Data Science and Machine Learning*, chapter Serialization in Object-Oriented Programming languages, pages 873–891. InTech.

Gumiel, M., da Mota, F. F., Rizzo, V. S., Sarquis, O., de Castro, D. P., Lima, M. M., Garcia, E. S., Carels, N., and Azambuja, P. (2015). Characterization of the microbiota in the guts of *Triatoma brasiliensis* and *Triatoma pseudomaculata* infected by *Trypanosoma cruzi* in natural conditions using culture independent methods. *Parasit Vectors*, 8:245.

Gumiel, M., de Mattos, D. P., Vieira, C. S., Moraes, C. S., Moreira, C. J. C., Gonzalez, M. S., Teixeira-Ferreira, A., Waghabi, M., Azambuja, P., and Carels, N. (2020). Proteome of the triatomine digestive tract: From catalytic to immune pathways; focusing on annexin expression. *Front Mol Biosciences*, 7:1–23.

Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Semin Cell Dev Biol*, 20:869–876.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model*, 52(7):1757–

1768.

Kanehisa, M. and Goto, D. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30.

Kindler, E. and Krivy, I. (2011). Object-Oriented Simulation of systems with sophisticated control. *Int J Gen Syst*, 40(3):313–343.

Lechuga, G. C., Souza-Silva, F., Sacramento, C. Q., Trugilho, M. R. O., Valente, R. H., Napoleão-Pego, P., Dias, S. S. G., Fintelman-Rodrigues, N., Temerozo, J. R., Carels, N., Alves, C. R., Pereira, M. C. S., Provance, D. W. J., Souza, T. M. L., and De Simone, S. G. (2021). SARS-CoV-2 proteins bind to hemoglobin and its metabolites. *Int J Mol Sci*, 22(16):9035.

Lima, C. R., Carels, N., Guimarães, A. C. R., Tuffery, P., and Derreumaux, P. (2016). *In silico* structural characterization of protein targets for drug development against *Trypanosoma cruzi*. *J Mol Model*, 22:244.

Lima, L. S., Gramacho, K., Pires, J. L., Clément, D., Lopes, U. V., Carels, N., Gesteira, A., Gaiotto, F. A., Cascardo, J. C. M., and Micheli, F. (2010). Development, characterization, validation, and mapping of SSRs derived from *Theobroma cacao* L. - *Moniliophthora perniciosa* interaction ESTs. *Tree Genet Genomes*, 6:663–676.

Motta, S. and Pappalardo, F. (2013). Mathematical modeling of biological systems. *Brief Bioinformatics*, 14(4):411–422.

NRCCFICB (2005). *Catalyzing Inquiry at the Interface of Computing and Biology*. NationalAcademiesPress.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42:D358–D363.

Otto, T. D., Guimarães, A. C., Degrave, W. M., and de Miranda, A. B. (2008). AnEnPi: identification and annotation of analogous enzymes. *BMC Bioinformatics*, 9:544.

Pires, J. G., da Silva, G. F., Weyssow, T., Conforte, A. J., Pagnoncelli, D., da Silva, F. A. B., and Carels, N. (2021). Galaxy and MEAN Stack to create a user-friendly workflow for the rational optimization of cancer chemotherapy. *Front Genetics*, 12:155–181.

Ponce de Leon, M., de Miranda, A. B., Alvarez-Valin, F., and Carels, N. (2014). The purine bias of coding sequences is determined by physicochemical constraints on proteins. *Bioinf Biol Insights*, 8:93–108.

Sgariglia, D., Conforte, A. J., Pedreira, C. E., Vidal, C. L. A., Gonçalves, C. F. R., Carels, N., and Silva, F. A. B. (2021). Data-driven modeling of breast cancer tumors using Boolean networks. *Front Big Data*, 4:1–13.

Sgariglia, D., Gonçalves, C. F. R., de Carvalho, L., Pedreira, C., Carels, N., and Silva, F. A. B. (2023). Optimizing therapeutic targets for breast cancer using Boolean network models. *bioRxiv*, doi:10.1101/2023.05.10.540187.

Shannon, C. E. A. (1948). A mathematical theory of communication. *Bell Syst Tech J*, 27:379–423.

Takahashi, M., Kryukov, K., and Saitou, N. (2009). Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics*, 93(6):525–533.

The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. *Nucleic Acids Res*, 36:D440–D444.

Tilli, T. M., Carels, N., Tuszynski, J. A., and Pasdar, M. (2016). Validation of a network-based strategy for the optimization of combinatorial target selection in breast cancer therapy: siRNA knockdown of network targets in MDA-MB-231 cells as an in vitro model for inhibition of tumor development. *OncoTarget*, 7:63189–63203.

Vieira, M., Gonçalves, C. F. R., Côrtese, A., Carels, N., and Silva, F. A. B. (2023). Statistical characterization of the dynamics of Glioblastoma Multiforme subtypes through parameters estimation of the epigenetic landscape. *bioRxiv*, doi:10.1101/2023.04.25.538198.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., R., Cummings, Le, D., Pon, A., Knox, C., and Wilson, M. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, 46(D1):D1074–D1082.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M. A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Res*, 35:D521–D526.

Zhang, Y., Yang, H., Turra, D., Zhou, S., Ayhan, D. H., DeIulio, G. A., Guo, L., Broz, K., Wiederhold, N., Coleman, J. J., Donnell, K. O., Youngster, I., McAdam, A. J., Savinov, S., Shea, T., Young, S., Zeng, Q., Rep, M., Pearlman, E., Schwartz, D. C., Pietro, A. D., Kistler, H. C., and Ma, L.-J. (2020). The genome of opportunistic fungal pathogen *Fusarium oxysporum* carries a unique set of lineage-specific chromosomes. *Commun Biol*, 3:50.