Software de Análise de Anotação Genômica – GANAS: Uma ferramenta para análise de anotações genômicas

Genomic Annotation Analysis Software – GANAS: A tool for Genomic Annotation Analysis

DOI:10.34117/bjdv6n11-363

Recebimento dos originais: 19/10/2020
Aceitação para publicação: 17/11/2020

Pablo Alessandro Barbosa Viana
Mestrando em Biotecnologia em Saúde e Medicina Investigativa
Instituto Gonçalo Muniz Fiocruz
Endereço: Rua Waldemar Falcão, 121, Candeal – Salvador - BA, 40296-710.
E-mail: pablo.alessandro@gmail.com

Elen Bethleen de Souza Carvalho
Doutorado em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: elenbeth@yahoo.com.br

Whendel Mesquita do Nascimento
Doutorando em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: whendelmesquita@ufam.edu.br

Roberto Alexandre Alves Barbosa Filho
Doutorando em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: robertoaabfilho@gmail.com

David Silva de Oliveira
Mestrando em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: robertoaabfilho@gmail.com

Enedina Nogueira Assunção
Doutorado em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: dinanog@yahoo.com.br

Spartaco Astolfi Filho
Doutorando em Biotecnologia
Universidade Federal do Amazonas
Endereço: Av. General Rodrigo Octavio Jordão Ramos, 1200 - Coroado I, Manaus - AM, 69067-005.
E-mail: spartaco.biotec@gmail.com

RESUMO
O desenvolvimento do sequenciamento de nova geração, impulsionou expressivamente a capacidade de geração de dados genômicos, aumentando assim a demanda por ferramentas automatizadas de análises genômicas. Após o sequenciamento e a montagem do genoma é realizada a anotação genômica, processo que permite a extração de dados relevantes das sequências geradas, destacando a identificação dos genes codificadores de proteínas. Com o objetivo inicial de facilitar a expansão da análise *off-line* dos genes de resistência a metais pesados da espécie *Enterobacter cloacae amazonensis* foi desenvolvido o programa Genomic Annotation Analysis Software - GANAS, que detalha visualmente os genes e subsistemas metabólicos anotados com a utilização da plataforma RAST. Os genes anotados foram analisados pelo software comparando os resultados em diversos organismos com conhecida capacidade de resistência a metais pesados. Foi verificado que esta cepa possui mais genes de resistência que os principais padrões encontrados na literatura, assim, evidenciando a eficiência da utilização do GANAS.

Palavras-Chave: Bioinformática, Genoma, Bactéria.

ABSTRACT
The development of new generation sequencing has significantly boosted the capacity to generate genomic data, thus increasing the demand to automated genomic analysis tools. After genome sequencing and assembly, genomic annotation is performed, a process that allows the extraction of relevant data from generated sequences, highlighting the identification of protein coding genes. In order to facilitate the expansion of offline analysis of genes for resistance to heavy metals of the species Enterobacter cloacae amazonensis, the software Genomic Annotation Analysis Software - GANAS was developed, which visually details the genes and metabolic subsystems annotated using the platform RAST. The annotated genes were analyzed by software comparing the results in several organisms with known resistance to heavy metals. It was found that strain has more resistance genes than the main patterns found in the literature, thus showing the efficiency of using GANAS.

Keywords: Bioinformatics, Genome, Bacteria.

1 INTRODUÇÃO

O advento das tecnologias de sequenciamento de DNA em conjunto com a bioinformática modificou radicalmente o cenário biológico científico, ajudando a caracterizar os mais diversos organismos. Atualmente, a Biologia Molecular e a Bioinformática possuem uma íntima relação, tornando difícil o desenvolvimento de estudos sem o conhecimento em ambas as áreas (Gauthier et al, 2018, Santana, et al. 2020).

Desde o início do século XXI, o desenvolvimento de novas tecnologias de sequenciamento de DNA, mais rápidas, baratas e precisas, impulsionou expressivamente o aumento da capacidade de

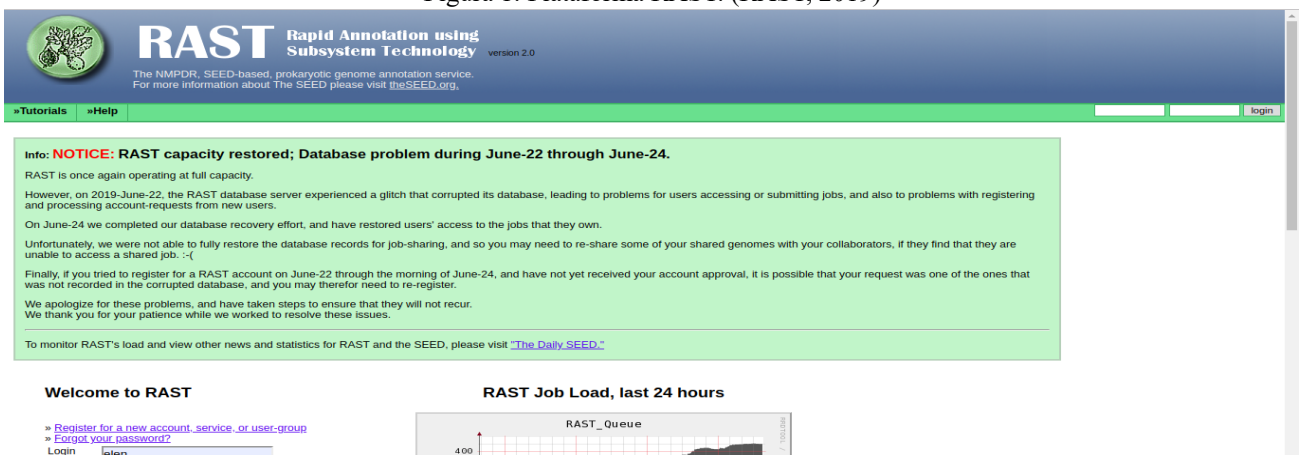
geração de dados genômicos. Porém, mesmo o sequenciamento do genoma de organismos simples como das bactérias, o volume de dados gerados é de alguns milhões de nucleotídeos, o que dificulta ou impossibilita seu tratamento manual, exigindo assim a utilização de ferramentas automatizadas (Lopes, 2015 e Gauthier et al, 2018).

Após a realização do sequenciamento, que comumente gera dados fragmentados, é realizada então a montagem do genoma, que por sua vez ordena esses diversos fragmentos em uma longa sequência de nucleotídeos. Para que seja possível extrair informações biológicas relevantes desta sequencia gerada, é realizado um processo chamado de anotação genômica. Este processo irá inferir estruturas e funções de elementos do DNA, identificando os genes codificadores de proteínas, RNA's não codificantes, regiões regulatórias, sequências repetitivas, entre outras estruturas (del Angel et al, 2018).

Apesar da anotação genômica envolver a caracterização de diversos elementos, maior atenção é direcionada à identificação dos genes codificadores de proteínas, em função principalmente das análises mais especializadas necessárias para caracterizá-los. Segundo o *National Center for Biotechnology Information* (2019), uma das ferramentas mais utilizadas no processo de anotação de genomas de organismos procariotos (bactérias e arqueas) é chamado de RAST (2019), *Rapid Annotation Subsystem Technology*. O RAST é um servidor de anotação online, gratuito e automatizado para genomas completos ou incompletos, de arqueas e bactérias.

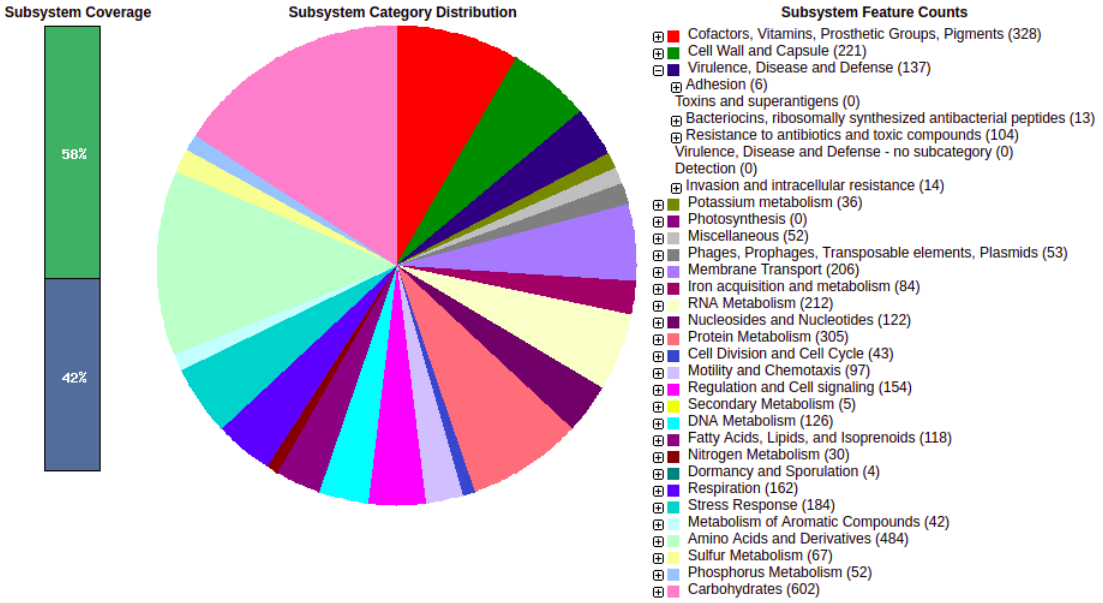
O RAST (**Figura 1**) foi inicialmente idealizado para o uso interno do *National Microbial Pathogen Data Resource* (NMPDR), mas rapidamente tornou-se globalmente utilizado. Neste servidor o usuário envia um arquivo em formato FASTA com um conjunto de *contigs* (longas sequências de DNA determinadas através do processo de sequenciamento) do genoma alvo e recebe acesso ao genoma anotado em um ambiente para análise e comparação com outras centenas de genomas preexistentes. (Aziz et al, 2008).

Figura 1: Plataforma RAST. (RAST, 2019)



Após determinar a função dos genes anotados, a plataforma RAST realiza uma reconstrução metabólica inicial, combinando genes do genoma adicionado pelo usuário a papéis funcionais de subsistemas predefinidos. Estes subsistemas estão agrupados em categorias refletindo os diversos módulos da maquinaria celular. Apesar de geralmente não ser possível estimar completamente os subsistemas de todos os genes, o conteúdo genômico que é conectado aos subsistemas provê uma detalhada estrutura para análise biológica, conforme evidenciado na **Figura 2** (Aziz et al, 2008 e Overbeek et al, 2013).

Figura 2: RAST: Genes conectados a subsistemas e sua distribuição em diferentes categorias. (RAST, 2019)



Apesar da plataforma RAST prover ferramentas próprias para análise e comparação de genomas, comumente análises mais específicas requerem a utilização de ferramentas externas que podem ser

encontradas também disponíveis publicamente pela comunidade científica e divulgada em artigos da área, ou que podem ser desenvolvidas pelo próprio pesquisador. Neste contexto destaca-se atualmente o uso da linguagem de programação Python, para o desenvolvimento de software para Bioinformática.

Por ser uma linguagem interpretativa de alto nível, ou seja, de fácil leitura e entendimento, o Python é uma linguagem poderosa e fácil de aprender, que segue um design que realça a legibilidade do código sobre a velocidade de execução, combinando uma sintaxe clara e concisa com os recursos poderosos de sua biblioteca padrão, por módulos e *frameworks* desenvolvidos por terceiros (*Python Software Foundation*, 2019).

A linguagem Python permite que se criem programas de acordo com vários paradigmas de programação, incluindo programação orientada a objetos, imperativa e funcional. Tornando ideal para escrever *scripts* (pequenos conjuntos de códigos em arquivos isolados) e para desenvolver rapidamente novas aplicações. (*Python Software Foundation*, 2019)

Buscando garantir a reprodutibilidade dos processos utilizados em Bioinformática, diversas medidas costumam ser utilizadas como padrão para desenvolvimento de métodos *in silico* buscando os princípios FAIR, *Findable, Accessible, Interoperable and Re-usable*, ou seja, encontrável, acessível, interoperável e reutilizável (Wilkinson, 2016). Entre estas medidas, está a utilização da plataforma Linux, garantindo mais estabilidade aos sistemas.

Além da comumente utilização da plataforma Linux para softwares de bioinformática, a utilização de métodos de conteinerização, que provem uma maneira de distribuir uniformemente os programas criados, a disponibilização dos projetos de desenvolvimento em repositórios abertos online como o github (*GitHub Inc*, 2019) e a utilização de ferramentas validadas pela comunidade como o RAST ou NCBI são estratégias que devem ser consideradas, para o desenvolvimento de novos softwares de bioinformática (del Angel et al, 2018).

Posto isso, o objetivo deste estudo foi desenvolver um programa em Python para analisar de modo *off-line* resultados das anotações genômicas obtidas através do servidor RAST, inicialmente para análise comparativa mais apurada das proteínas envolvidas no processo de biorremediação de metais pesados da bactéria *Enterobacter cloacae amazonensis* (Astolfi, 2018).

2 MATERIAL E MÉTODOS

O programa foi desenvolvido em Python, utilizando as bibliotecas **csv** (para ler arquivos .gff e .tsv), **re** (para trabalhar com expressões regulares) e **PyQt5** (para criar uma interface gráfica com janelas como um programa de computador comum, facilitando a sua utilização). (*Python Software*

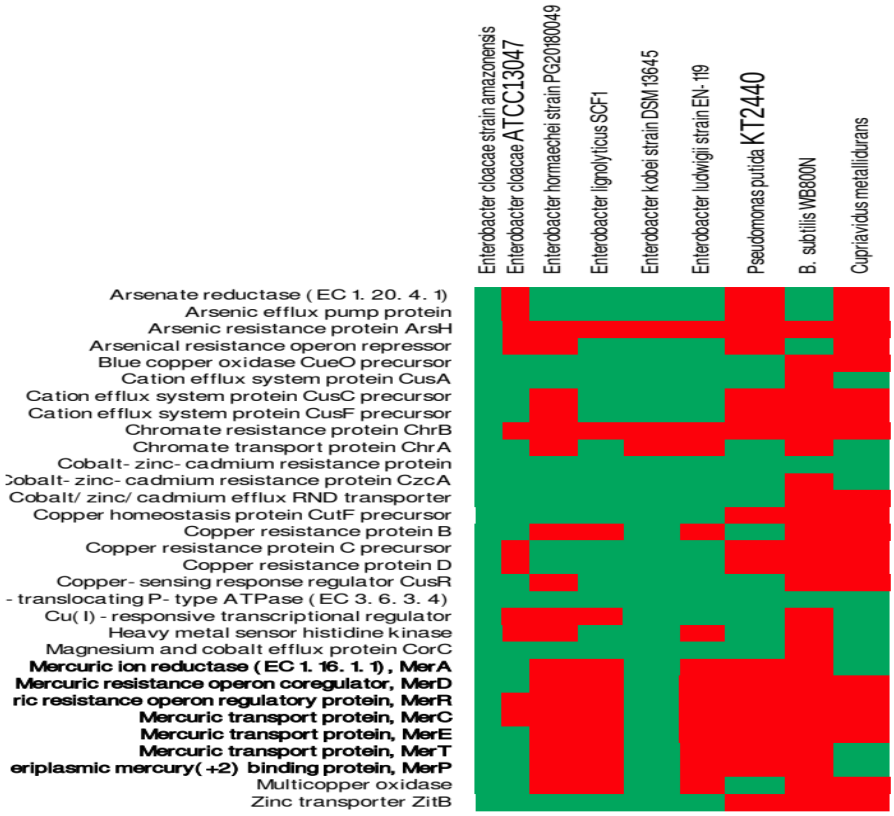
Foundation, 2019 e *The Qt Company*, 2019).

Para a validação do GANAs, a partir de dados parciais do genoma da *Enterobacter cloacae* amazonensis (Astolfi, 2018) e dos genomas de *Enterobacter hormaechei* strain PG20180049, *Enterobacter lignolyticus* SCF1, *Enterobacter kobei* strain DSM 13645. *Enterobacter ludwigii* strain EN-119, *Pseudomonas putida* KT2440, *B. subtilis* WB800N, *Cupriavidus metallidurans*, e do genoma de referência de *Enterobacter cloacae* ATCC 13047 (NC_014121.1), foram submetidos *contigs* na plataforma RAST realizada a anotação dos genes encontrados, para a validação do GANAS.

3 RESULTADOS E DISCUSSÃO

Neste estudo foi desenvolvido um programa capaz de carregar arquivos tanto arquivos**.gff** quanto**.tsv** concatenando suas informações de modo a estender as informações que podem ser utilizadas pelo usuário. Apesar do RAST prover informação de subsistemas e categorias de cada gene determinado, conforme a **Figura 2**, os dados só podem ser exportados separadamente, a informação dos genes e suas respectivas proteínas (formato**.gff**) e as categorias e subsistemas de cada gene (formato**.tsv**). O que torna laborioso o trabalho de filtrar mais apuradamente os genes que serão analisados, como separar informações dos genes de todas as proteínas de um determinado subsistema. Observando a tabela presente na **Figura 3,** verifica-se a grande quantidade de genes de resistência a metais pesados encontrados na *Enterobacter cloacae* amazonensis.

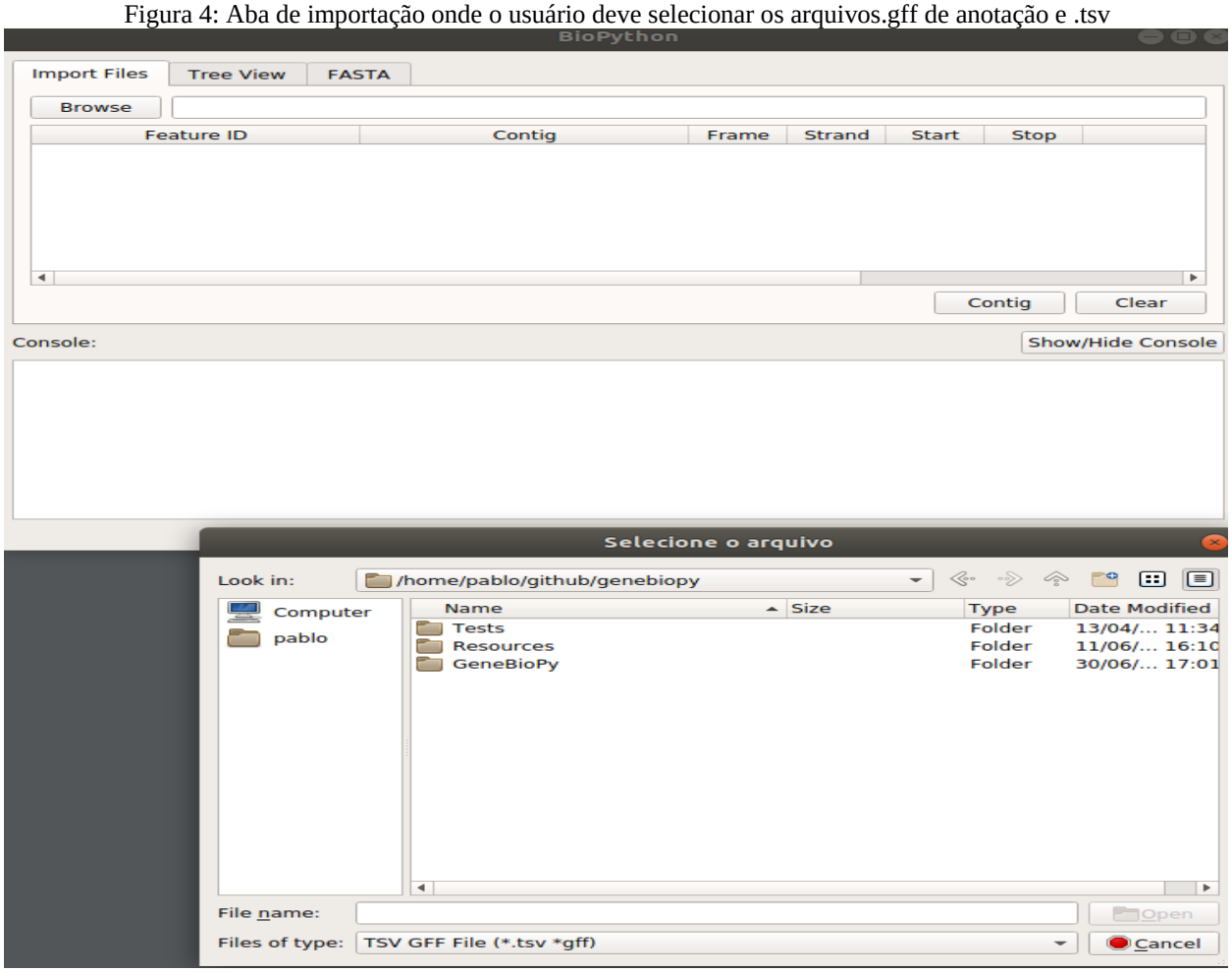
Figura 3: Tabela listando todos os genes de resistência a metais pesados. encontrados nos organismos de estudo, onde a célula marcada em verde representa a existência deste gene neste organismo e as células em vermelho representam a ausência do gene. Os números ao final da tabela representam o somatório de quantos dos genes listados este organismo possui.



A partir da análise da tabela é possível verificar que esta espécie possui mais genes de resistência aos mais variados tipos de metais pesados, de arsênio, cádmio e zinco, à mercúrio, cobre e cobalto. Esta identificação tornou-se mais fácil com a utilização da ferramenta desenvolvida, que também pode ser usada para explorar os outros atributos da anotação genômica, como a posição inicial e final dos genes, e utilizar esta informação para obter a sequência de DNA, combinando com o uso de outras ferramentas, como o BLAST (Basic Local Alignment Search Tool), para obter ainda mais informações e realizar análises mais apuradas dos genes encontrados e do potencial deste organismo.

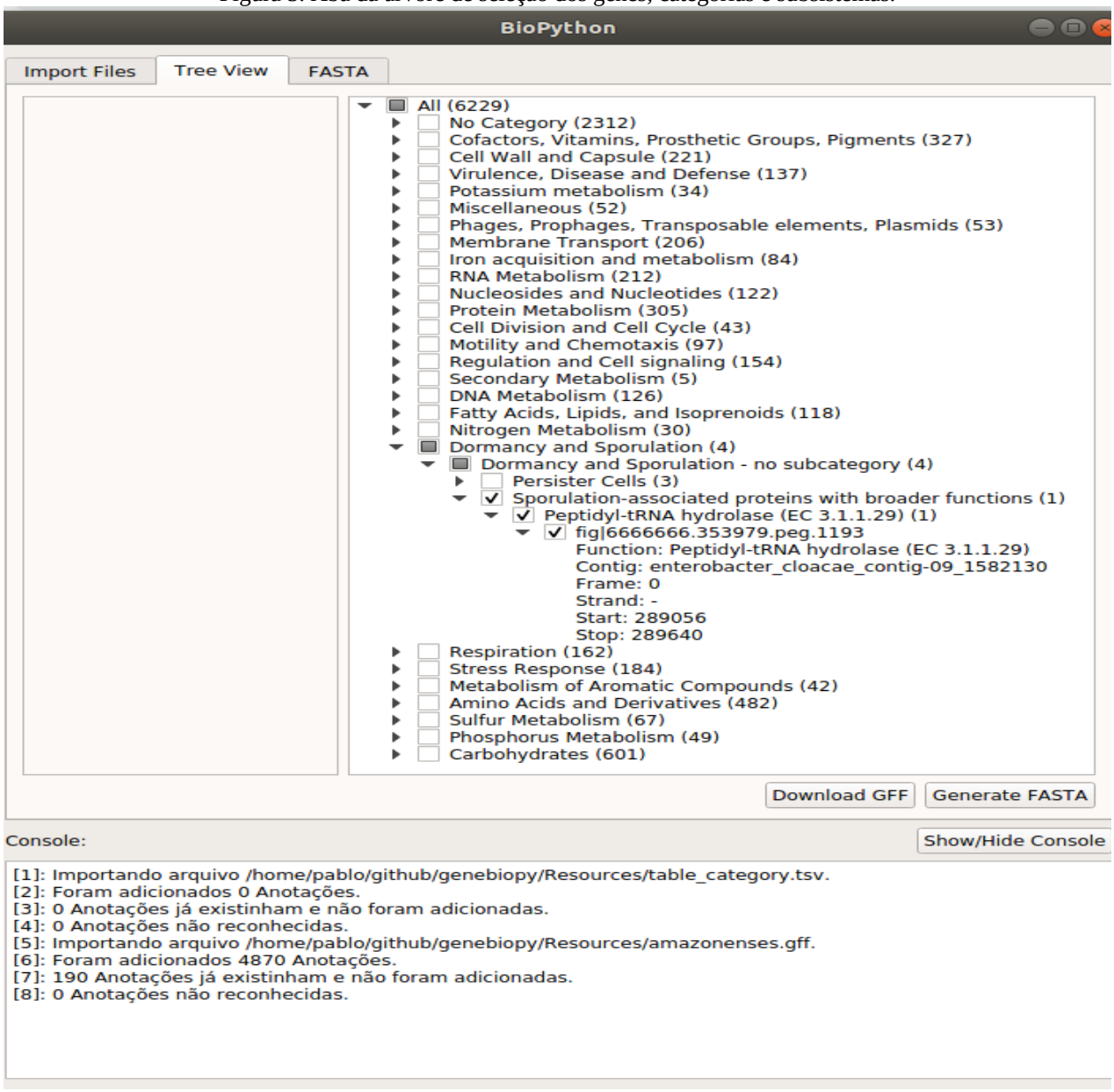
O programa desenvolvido encontra-se disponível publicamente no site https://github.com/biotecufam/genebiopy. Para executá-lo é necessário ter a plataforma Python 3.7 devidamente instalada no computador conjuntamente com o framework PyQt5.

Figura 4: Aba de importação onde o usuário deve selecionar os arquivos.gff de anotação e .tsv



Ao iniciar o programa o usuário é apresentado à tela de importação de arquivos conforme **Figura 4**, onde é possível carregar arquivos .gff ou .tsv, gerando uma estrutura de árvore similar à encontrada no RAST (**Figura 2**), com cada categoria, subsistemas e funções, porém agora acrescida de informações do gene específico e com uma caixa de seleção possibilitando ao usuário filtrar o que achar necessário, conforme figura 5.

Figura 5. Aba da árvore de seleção dos genes, categorias e subsistemas.



Após a seleção é permitido ao usuário fazer a exportação dos genes selecionados no formato **.gff**, para que estes possam ser carregados em outro programa ou utilizados para realizar alguma análise.

4 CONCLUSÃO

Concluímos que a utilização do software Genomic Annotation Analysis Software – GANAS, torna menos laborioso o trabalho de filtrar mais apuradamente os genes que serão analisados, assim como separar informações dos genes de todas as proteínas de um determinado subsistema, facilitando a análise de dados complexos.

REFERÊNCIAS

del ANGEL, V. D.; et al. Ten steps to get started in Genome Assembly and Annotation. F1000Reserach 2018, 7(ELIXIR):148. 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850084.1/.

ASTOLFI, M. C. T.; CARVALHO, E. B. de S.; de BARROS, A. M.; PINTO, M. V.; de LACERDA, L. B.; NOGUEIRA, V. B.; LOPES, E. F.; ASTOLFI-FILHO, S. Draft Genome Sequence of the Novel Enterobacter cloacae Strain amazonensis, a Highly Heavy Metal-Resistant Bacterium from a Contaminated Stream in Amazonas, Brazil. Genome Announcements, 2018, Vol. 6, Issue 22. American Society for Microbiology. 2018. https://mra.asm.org/content/6/22/e00450-18.short.

AZIZ, R. K.; et al. The RAST Server: Rapid Annotation Subsystems Technology. BMC Genomics 2008, 9:75. BioMed Central. 2008. https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-75?report=reader.

GAUTHIER, J.; VINCENT, A. T.; CHARETTE, S. J.; DEROME, N. A brief history of bioinformatics. Briefings in Bioinformatics, 2018, 1-16. Oxford Press. 2018. https://academic.oup.com/bib/article/20/6/1981/5066445.

GITHUB INC. GitHub. Disponivel em: https://www.github.com. Acessado em: 26 de junho de 2019.
LOPES, R. B. M. Montagem e análise do genoma parcial de Bacillus thuringiensis BAC3151 endofítico das folhas do feijoeiro comum (Phaseolus vulgaris). Dissertação de Mestrado - Universidade Federal de Viçosa. 2015. https://www.locus.ufv.br/bitstream/handle/123456789/6497/texto%20completo.pdf?sequence=1

NCBI. National Center for Biotechnology Information. Disponível em: https://www.ncbi.nlm.nih.gov. Acessado em: 26 de junho de 2019.

OVERBEEK, R.; et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 2014, Vol. 42, Database issue. Oxford Press. 2013. https://academic.oup.com/nar/article/42/D1/D206/1062536

PYTHON SOFTWARE FOUNDATION. Python programming language. Disponível em: https://www.python.org. Acessado em: 26 de junho de 2019.

THE QT COMPANY. Qt. Disponivel em: https://www.qt.io. Acessado em: 26 de junho de 2019.
RAST. Rapid Annotation using Subsystem Technology. Disponível em: https://rast.theseed.org. Acessado em: 26 de junho de 2019.

SANTANA, R.S.; COSTA, E.A; MELO, S.C.O; PUNGARTNIK, C. Characterization of the enzyme Oxidase Alternative of the fungus Moniliophthora perniciosa with the aid of bioinformatics tools. Braz. J. of Develop., Curitiba, v. 6, n. 10, p. 81420-81430, oct. 2020.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018. 2016. https://www.nature.com/articles/sdata201618%22