Article │ Received 22 May 2024; Accepted 24 September 2024; Published 29 September 2024
https://doi.org/10.55092/bi20240001

Pannotator integrated with Medpipe provides immunological and subcellular location features using a microservice

Rafael Gonçalves[1] and Anderson Santos[1,*]

[1] Faculty of Computing, Federal University of Uberlândia, Uberlândia, Brazil

* Correspondence author; E-mail: santosardr@ufu.br.

Abstract: Bacterial and archaea genome sequencing and assembly are trivial tasks nowadays. After assembling contigs and scaffolds from a genome, the subsequent step is annotation. An annotation evidencing the expected features, like rRNA, tRNA, and CDS, is a signal of the quality of our sequencing and assembly. Different techniques to obtain and reproduce DNA samples, as well as sequencing and assembly of genomes, can impact the quality of a genome's expected features. The Pannotator tool was conceived as an aid annotation tool focusing on the differences between assembling and its reference genome. Some of the key features for bacterial genome annotations are the subcellular location and immunological potential of a CDS. Instead of reimplementing the prediction of these features in Pannotator, we leveraged the capabilities of our microservice to provide them. In the end, Medpipe software was not modified, and Pannotator underwent minor changes to incorporate the subcellular location and immunological potential of all exported proteins annotated by the tool. Moreover, our Medpipe microservice can also be incorporated into other software. The Medpipe microservice is open to anyone, not only to our Pannotator tool. The successful integration of Medpipe to Pannotator, powered by the Medpipe microservice, offers a powerful approach to advanced genomic analysis. The Medpipe microservice, built on Kotlin with the Spring Boot framework, is instrumental in the automation of Medpipe processing. It achieves this using REST endpoints, such as the execution of Medpipe in an asynchronous manner, status retrieval, and prediction generation, which enhance the modularity and scalability of the microservice. The availability of endpoint documentation, detailed request examples, and logs make our microservice user-friendly. The results of this integration demonstrate the value of the information provided by Medpipe, enriching genomic annotation with additional details, such as the density of mature epitopes (MED) and protein subcellular location classification. The Pannotator has evolved beyond basic function annotation and now provides data on immunological potential, structure, and subcellular location after being integrated with our microservice. The Medpipe microservice is available at https://github.com/santosardr/medpipe-ms.git.

Keywords: genome; annotation; medpipe; pannotator; microservices; protein; subcellular location; mature epitope density

1. Introduction

Bioinformatics is an area of research that integrates computer science, statistics, mathematics, and biology. She has been instrumental in advancing research genomics, providing tools and methodologies for data processing and analysis [1]. In this context, sequence alignment emerged as a fundamental pillar, providing a deep insight into evolutionary relationships and functional disorders between genomes. Biological sequence alignments are tools that, in addition to being used for the analysis of conserved regions and regions that have suffered mutations in homologous sequences, also serve as a starting point for other applications in Computational Biology, such as the study of secondary protein structures and the construction of phylogenetic trees [2]. The alignment of sequences not only unravels the intricate tapestry of genomic evolution but also serves as a crucial starting point for further investigations. A vital application of this process lies in genomic annotation, where the identification of functional elements, such as genes and regulatory regions, constitutes the essence of the decoding of the genetic code. According to [3], the annotation process is crucial for developing methodologies based on the analysis of genetic material, such as pan-genomics and taxonomies. Genomic annotation is a complex process that requires the identification and cataloging of functional elements in a genome, and the large amount of data generated by sequencing techniques makes this task even more complicated. In this context, the Pannotator (pannotator.facom.ufu.br) is a tool designed to create an automatic annotation from a reference genome. This tool has been developed with the aim of minimizing the workload required in the preparation of and correction of various annotations during the execution of a pan-genome [4]. However, the search for further improvements and predictions has led to the introduction of Mature Epitope Density (Medpipe). Medpipe (medpipe.facom.ufu.br) is a bioinformatics pipeline designed to predict epitope density by mature portions of proteins [5]. Our web pages include Medpipe and Pannotator. In Pannotator, proteins are used to perform the Basic Local Alignment Search Tool (BLAST), which compares proteins from reference genomes.

Data on subcellular potential localization (cytoplasm, cell wall, surface exposed and secreted) of proteins generated by Medpipe may be useful to a Pannotator user who is annotating the genome of a bacterium. Until this work, Pannotator users could count on integration with Medpipe to request an annotation service and incorporate it into the annotation of a genome by the Pannotator. Such integration has the potential to reduce the work of a researcher to run the Pannotator and Medpipe separately. In addition, the service offered by Medpipe is unique, and there is no other tool that provides the same service on the Internet. Since Medpipe is a non-free software, open-sourcing is not an alternative. An elegant solution is a communication service between Medpipe and any other biological sequence analysis software. The implementation of a microservice executing and returning the results will meet the expectations of providing additional data for the Pannotator and any other software.

The term “Microservices Architecture” has emerged in recent years to describe a particular way of designing software applications as sets of independently implemented services. While there is no precise definition of this style of architecture, there are certain common characteristics surrounding the organization of business capability, automated deployment, intelligence in terminals, and decentralized control of languages and data [6]. One of the characteristics of using Microservices Architecture is being able to design a set of independent services, each responsible for a specific part of the functionality, built around business capabilities, which means that they are designed to meet the needs of a particular business. These services deployed independently can be upgraded or retired without affecting the other services. Management is decentralized, which means that everyone is responsible, in addition to being able to write in different programming languages and use different data storage technologies, which gives organizations more flexibility [6].

According to [7], microservices offer benefits such as better scalability, faster deployment, flexibility, resiliency, smaller development teams, modularity, reliability, and reusability. However, there are drawbacks and present challenges in their use, as specified by Velepucha and Flores [8]. This new architecture has a high learning curve, requires little experience in team development, has network overhead, is more complex, and duplicates data.

A scalable, high-performance microservice is about more than just the ability to handle a large volume of tasks or concurrent requests. The essence of this lies in its efficiency in performing operations. It can process tasks quickly and effectively, use resources optimally, and maintain high performance even under heavy loads without compromising service quality. It is also prepared for future demand growth by adapting and scaling accordingly [9]. The flexibility to make modifications to a single service and deploy it in an independently optimized software development and delivery allows the rapid implementation of new features and bug fixes. That autonomy also makes it easier to isolate faults if they occur, restricting the problem to the service in question and simplifying its resolution. In addition, the microservices architecture Enables the reversal of changes quickly and minimizes the risks and negative impacts [10].

Many companies use microservices in their solutions; an example is Netflix, which has developed a microservices platform centered on media workflows, aiming to increase flexibility and speed of development [11]. Another example is Amazon, which faced scalability and productivity issues due to frequent updates, and projects that faced the need to refactor their system and took the microservices approach. The division of monolithic applications into small independent services has enabled the company to respond better to the requirements and make rapid and specific changes [12].

Based on the microservices architecture, we designed the Medμσ that makes use of the Medpipe pipeline. This approach ensures the rapid availability of new features to users. Although the present work shows the integration between Pannotator and Medpipe through the Medpipe microservice (Medμσ), it is essential to point out that the microservice developed is not specific to the use of Pannotator, which can be used by any software or analysis pipeline of biological sequences through the availability of the endpoints on the internet.

The main objective of this work was to create a microservice with the name of Medμσ, which executes Medpipe through an Application Programming Interface (API) following the architecture style Representational State Transfer (REST). We also integrated these APIs with Pannotator, providing a platform accessible to researchers and professionals working with genome assembly and genomic sequence analysis. Specifically, our objectives included describing the development of the Medμσ and integrating the APIs developed in the microservice into Pannotator.

2. State-of-the-art related research

In this section, we discuss works with objectives like the proposal of the Pannotator and Medpipe.

2.1. Review of the pannotator literature

The multiplex capability and high throughput of sequencing instruments have made the complete sequencing of the bacterial genome routine. Over the past ten years, several automated genome annotation tools have been made available as open-source software or on publicly accessible pages on the internet. In the following paragraphs, we will concisely address the main features of these tools.

The Prokka is a command-line tool implemented in Perl that allows annotation completion of a preliminary bacterial genome in about 10 minutes on a computer desktop, producing standards-compliant output files for further analysis or visualization in genomic browsers, outperforming web and email-based systems that are unsuitable for sensitive data or integration into computational pipelines [13].

BG7 is an open-source tool based on an annotation paradigm protein-centric gene genomes, designed specifically for bacterial genomes sequenced with next-generation sequencing (NGS) technologies, considering the peculiarities of bacterial genomes (absence of introns and scarcity of sequences non-coding proteins) and NGS technologies, being able to deal with errors and annotate highly fragmented genomes or mixed sequences of several genomes (such as those obtained by metagenomics samples). It has been designed for scalability by using a cloud computing infrastructure based on Amazon Web Services (AWS) [14].

The RAST tool kit (RASTtk) is a modular version of the annotation engine RAST that allows researchers to build custom annotation pipelines, choose software to identify and annotate genomic traits, add features custom to an annotation job, accommodate batch genome submissions, and customize annotation protocols for batch submissions. RASTtk marked the first major software restructuring of RAST since its inception in 2008 [15].

The Protein Sequence Annotation Tool (PSAT) is a meta-platform based on the Web for integrated, high-throughput analysis of genomic sequences, demonstrating its usefulness in annotating the gene products of predicted peptides of *Herbaspirillum sp. strain RV1423*, import the results into the EC2KEGG, and use the resulting functional comparisons to identify a putative catabolic pathway, highlighting the potential in a genome with limited annotation [16]. Genix is a web-based bacterial genome annotation platform, which stands out for providing results closer to the reference annotation, with a lower number of false-positive

proteins and non-annotated functional proteins, being able to enhance the accuracy of bacterial genome annotation steps and provide high-quality results [17].

Sma3s is an accurate computational tool for automatic protein annotation with useful functionalities for fundamental and applied science. It provides functional categories and requires low computational resources, allowing complete annotation of proteomes and transcriptomes in about 24 hours on a personal computer [18].

ProGenomes is a comprehensive database containing high-quality genomic information from a wide variety of microorganisms. It provides access to bacterial, archaeal, and eukaryotic genomes, along with metagenomes and plasmids. ProGenomes is a valuable tool for comparative genomics studies, evolution and microbial research, and new species research [19].

DFAST, a genome annotation pipeline for prokaryotes, also assists in sending data to the public sequence database, with emphasis on its ability to annotate a typical-sized bacterial genome in less than 5 minutes and its integration with the DNA Data Bank of Japan (DDBJ) [20].

EuGene is a gene search tool that can be used in the genomes of prokaryotes and eukaryotes. It uses statistical information, similarities with genes and known proteins, and structured data in GFF3 format to predict the unit transcription of the main genes in the genome and perform functional annotations. This tool can deal with complex genomes with repeating regions and transposable elements and can be configured as ab initio, similarity-based, or hybrid, depending on the sources of information used [21].

DescribePROT is a database that contains 13 descriptors predicted in amino acid level for protein structure and function, encompassing 83 proteomes complete model organisms and including 7.8 billion predictions for 600 million amino acids in 1.4 million proteins, with the possibility of searching for amino acid sequences and UniProt accession numbers [22].

µProteInS is a proteogenomic pipeline implemented in Python 3.8. It combines genomics, transcriptomics, and proteomics to identify microproteins in bacteria, overcoming the limitations of traditional approaches and enabling the identification of Small ORFs (smORFs) with overlapping genes, leaderless transcripts, and sequences not preserved. µProteInS is distributed as open-source software [23].

2.2. Review of the literature on medpipe

We present here a review of the literature on software and web servers that have been available for the last ten years to find candidate proteins for vaccines or diagnostic targets against pathogenic bacteria. Surprisingly, studies have yet to be conducted on this theme.

Jenner-Predict is a web server that uses a knowledge-based approach to bacterial pathogenesis to predict protein-based vaccine candidates (PVCs) from proteomes of bacterial pathogens. The web server considers domains of different classes of proteins involved in host-pathogen interactions and pathogenesis, including adhesins, virulence, invasins, porins, flagellin, toxins, and others. In addition, Jenner-Predict evaluates the potential immunogenicity of PVCs, comparing them to known epitopes and considering the absence of autoimmunity and conservation in different strains. The server has demonstrated high accuracy in predicting known PVCs and overcame existing methods such as NERVE, Vaxign, and VaxiJen [24].

Another site, VacTarBac, is a web server that uses an immunoinformatic approach to identify vaccine candidates based on epitopes against 14 species of pathogenic bacteria. The server utilizes a comprehensive analysis of target proteins, including virulence factors and essential genes, to predict epitopes with the potential to stimulate different components of the immune system. In addition, VacTarBac removed self-recognized epitopes to prevent unwanted immune responses. The analysis revealed 21 proteins from 5 bacterial species as targets of promising vaccines. The server also identifies B-cell epitopes, T cells, and MHC-II ligands, as well as adjuvants, resulting in a total of 252 unique epitopes. VacTarBac features visualization to assist users in identifying the best vaccine candidates in an antigenic sequence [25].

The Integrative Vaccine Investigation and Online Information Network (VIOLIN) is a database and vaccine research analysis system that curates, stores, analyzes, and integrates diverse vaccine-related research data. Since its first publication in 2008, VIOLIN has undergone significant updates. It currently includes more than 3240 vaccines for 192 infectious diseases and eight non-infectious diseases. Within VIOLIN, there are more than ten independent programs. As an example of programs, we can cite Teeth, such as Protegen, which stores antigenic proteins that have been proven to be valid for vaccine development, and VirmugenDB, which annotates virulence factor genes that can be mutated to generate attenuated vaccines successfully. The VIOLIN also includes Vaxign, the first vaccine candidate prediction program based on reverse vaccinology, and other components of vaccines, such as adjuvants (Vaxjo) and DNA vaccine plasmids (DNAVaxDB). In addition, VIOLIN has databases of licensed human vaccines (Huvax) and veterinary vaccines (Vevax). The Ontology of Vaccines is applied to standardize and integrate the different data in VIOLIN. The VIOLIN also hosts the Vaccine Adverse Event Ontology (OVAE), which represents adverse events associated with licensed human vaccines [26].

A non-web-based alternative is TiD, a stand-alone application that identifies potential targets for drug development. It uses the premise that a protein must be essential for the survival of the pathogen and not homologous to the host to qualify as a target. TiD removes paralogous proteins, selects essential organisms, and excludes those that are homologous to host organisms. Targets are classified as known, new, or virulent. Users can perform road analysis metabolic, protein interactions, and other functionalities through integrated web servers. Targets identified by TiD for Listeria monocytogenes, Bacillus anthracis, and Pseudomonas aeruginosa showed overlap with previous studies. TiD is a useful tool for the rational development of medicines, as it analyzes targets in a bacterial proteome in about two hours [26].

2.3. Annotation and pathogenicity integration

In addition to providing functional annotation, some tools are also available to provide additional annotation data regarding the type of protein export, resistance to antibiotics, structure, and immune capacity of proteins. For example, MacSyFinder is a bioinformatics tool that allows the search for genetic systems in whole genomes. It utilizes a standards-based approach and domain profiles to identify and annotate secretion systems, antibiotic resistance

systems, and other genetic systems in different organisms. The MacSyFinder is distributed as an open-source software [27].

In the chapter "Antigen discovery in bacterial pan-proteomes", an in-silico methodology is described which integrates pan-genomic, immunoinformatic, structural, and evolutionary approaches to screening for potential antigens in each bacterial species, aiming at the development of broadly protective vaccines and avoiding specific immunity to alleles, in addition to allowing the development of diagnostic assays [28].

The purpose of our work was to follow these software examples, integrating the Pannotator with Medpipe and providing, along with functional annotation, data that helps to find proteins in a genome with potential for vaccine production and testing of pathogen diagnostics. However, our proposal went further when it implemented a protein pathogenicity annotation microservice that any other genomic annotation software can use. Our software, rather than a competitor, is functional annotation software, which proposes to be a partner of the most functional annotation capabilities that currently exist when offering a microservice that can be consumed democratically.

3. Methods

In this section, we will detail how the Medμσ was conceived, designed, and built, as well as the technologies used and the implementation of the crucial endpoints.

3.1. Microservices architecture

Microservices are built around businesses and can be deployed independently [6]. As shown in Figure 1, each microservice can be responsible for a specific biological analysis or query and function independently without affecting the others. In this example, MS-1 could be a service for running genetic analysis tools on a DNA sequence, MS-2 is a service that performs protein structure prediction using a protein sequence, and MS-3 is a service for searching for specific gene sequences in a genome. The three microservices are independent, so it makes it easy to make any changes to any of them. If the need arises to add a new genetic analysis tool, only MS-1 will be affected without any side effects from MS-2 and MS-3. As shown in Figure 1, each microservice can be responsible for a specific business and function independently without affecting the others. In this example, MS-1 could be a registration service for employees of a company, MS-2 is a service that performs tax calculations for the company, and MS-3 is a service for supplier registration of the company. All three microservices are independent, so it makes it easy to make any changes to any of them. If the company needs to add a new tax calculation rule, only MS-2 will be affected without any side effects from MS-1 and MS-3. In the case of Medμσ, which was built for this work, we followed the same logic. If it is necessary to run or search for results from a bioinformatics pipeline other than Medpipe, another microservice should be built. The Medμσ was constructed only for the execution and search of Medpipe results. Any new functionality that is not part of Medpipe should not be included in the Medμσ, from which one should think about the construction of a new microservice.

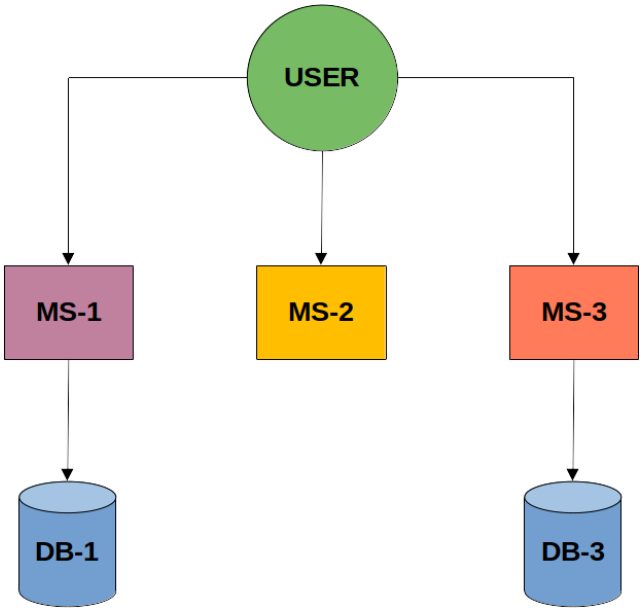

Figure 1. Example of microservices architecture.

We followed the same logic in the case of Medμσ, which was built for this work. Medμσ was constructed only for the execution and search of Medpipe results. Any new functionality that is not part of Medpipe should not be included in Medμσ, from which one should think about the construction of a new microservice. Figure 2 depicts the Medμσ schema.

The supplementary material documenting and instructing how to use Medμσ in your bioinformatic tool is accessible in the README file from the GitHub repository at https://github.com/santosardr/medpipe-ms.git.

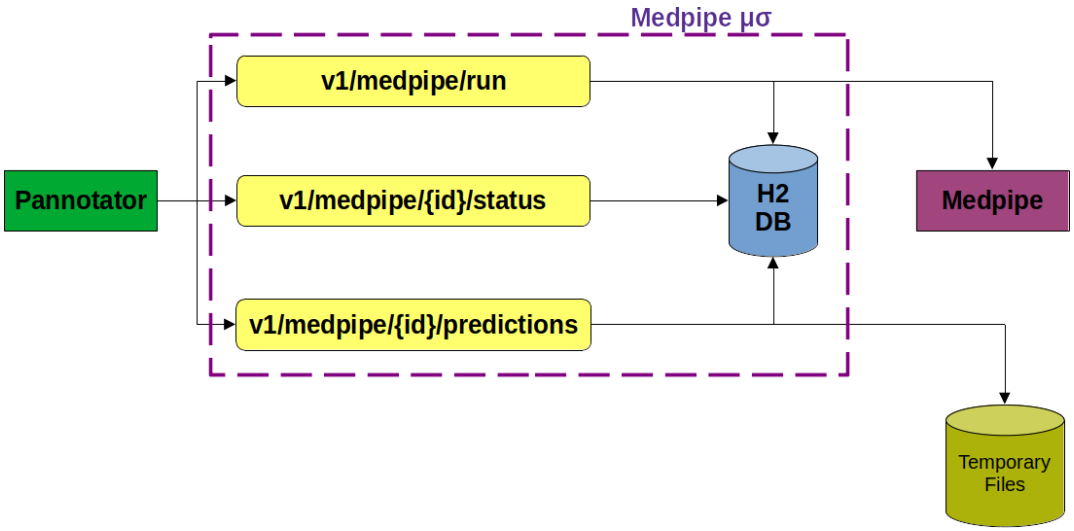

Figure 2. Medpipe microservice architecture. The Pannotator integrates with the Medpipe microservice using the endpoints. Pannotator and Medpipe are basically bash scripts that can be run from the command line or through their web interfaces. With the development of our microservice, Pannotator now runs Medpipe through an API call that is exposed on the internet. The execution of Medpipe is done by the /v1/medpipe/run route and has been included in the Pannotator after updating the target fasta file.

3.2. Benefits of microservice architecture in genomic annotation

Microservice architecture offers significant advantages for the genomic annotation process, enhancing flexibility, scalability, and integration capabilities. Below are specific benefits derived from our manuscript.

- Independent service functionality: Microservices can function independently, allowing specific biological analyses to be executed without impacting other services. For instance, separate microservices can handle tasks such as genetic analysis, protein structure prediction, and gene sequence searching [7]. This independence facilitates quick updates and modifications to individual services without disrupting the overall system, enhancing the genomic annotation workflow.
- Modular design: The modular nature of microservices enables easier management and development of genomic annotation tools. Each microservice can be developed, deployed, and maintained by smaller teams focused on specific functionalities [3]. This results in improved development speed and allows for the integration of new features as they become available, such as advanced annotation capabilities.
- Scalability and flexibility: Microservices provide better scalability by allowing organizations to scale specific services independently based on demand. This is particularly useful in genomic annotation, where data loads can vary significantly [3]. Organizations can adapt their architectures according to project needs, enhancing flexibility in resource allocation and service deployment.
- Enhanced collaboration: Microservices promote cooperation among different tools and platforms. For example, the Medpipe microservice can be utilized by various genomic analysis applications beyond Pannotator, allowing for a broader range of functionalities to be integrated seamlessly [4,7]. This encourages the sharing and reuse of services across different projects, fostering innovation and reducing redundancy.
- Improved deployment speed: Microservices support faster deployment cycles. By deploying individual components, teams can implement updates and new features more rapidly, which is crucial in the fast-evolving field of genomics [3]. This agility leads to quicker iterations of genomic analysis methods, ensuring that researchers have access to the latest tools and data.
- Resilience and fault isolation: The architecture's design helps isolate faults in specific services, minimizing the impact on the overall system. If one microservice fails, it does not necessarily bring down the entire genomic annotation process [3]. This resilience contributes to more reliable genomic analysis outcomes and less downtime in data processing.
- Comprehensive documentation and user experience: Microservices often come with extensive documentation, making it easier for users to understand and utilize the APIs effectively. The Medpipe microservice provides detailed endpoint documentation, enhancing user experience and accessibility for researchers [1,4]. This user-centric approach helps streamline the integration of genomic tools, making sophisticated analyses more approachable.

- Automation of processes: Microservices facilitate the automation of various genomic annotation tasks. For example, the Medpipe microservice automates the execution of analyses through REST APIs, allowing for asynchronous operations that enhance processing efficiency [1,4,7]. Automation reduces the manual effort required in genomic analyses, freeing up resources for more complex tasks.

In summary, microservice architecture significantly enhances the genomic annotation process through its independent functionality, modular design, scalability, and resilience. The ability to integrate seamlessly with various tools and the promotion of rapid deployment cycles further solidify its importance in advancing genomic research and applications.

3.3. Programming language

Kotlin was chosen for microservice development. Created by JetBrains, Kotlin is a modern programming language that is secure, expressive, and interoperable with Java [29]. Compared to Java, Kotlin minimizes the need for boilerplate code, resulting in clearer, more maintainable code.

3.4. Spring boot

Spring Boot is a Java application development framework that is based on the principle of "only what is necessary". It provides a few features and tools needed to build Java applications, but it only offers essential features. Right makes Spring Boot a lightweight and easy-to-use platform [30]. One of Spring Boot's outstanding features is its "opinion on configuration" approach. This approach provides sensible default settings for many application aspects, allowing developers to focus more on business logic than on complex configurations. In addition, Spring Boot offers an integrated system for building and managing dependencies, which simplifies the management of the project's required libraries.

3.5. Maven

Maven is a built-in automation and project management tool based on an artifact model. An artifact model is a framework that defines the artifacts a project can have, such as source code, libraries, configuration files, and other files. Maven uses the Artifact template to automate the build, test, and packaging processes and project deployment [31].

3.6. H2 database

The H2 database, a relational database written in Java, was chosen to store information related to the integration processes. It can be Run in client-server mode or embedded mode. In client-server mode, the database runs on a separate server from the Java application. In inline mode, the database runs in the same process as the Java application figure.

3.7. The integration process between the pannotator and medpipe

The integration of Pannotator with Medpipe is designed to enhance genomic analysis capabilities by providing detailed functional annotations and immunological insights. Pannotator creates automatic annotations from reference genomes, while Medpipe predicts epitope density, enriching the annotation process. These are the key components of the Integration:

- Microservice architecture: the integration utilizes a microservice called Medμσ, which executes Medpipe through a RESTful API. This architecture enhances modularity and scalability, allowing for seamless communication between the two tools [1,4].
- API endpoints: Medpipe is accessed via specific endpoints, such as the /v1/medpipe/run, which allows Pannotator to perform annotations efficiently. These endpoints facilitate asynchronous execution and status retrieval [1,11].
- Functional enhancements: with the integration, the Pannotator can now provide data on proteins' Subcellular location, Immunological potential, Mature Epitope Density (MED), and Classification in relation to Gene Ontology (GO) [1,14].
- User-Friendly features: the microservice offers detailed documentation, example requests, and logs, making it accessible for researchers to utilize genomic data effectively [1].
- No modification required for Medpipe: the integration did not necessitate changes to the Medpipe software. Only minor adjustments were made to Pannotator to incorporate additional features [1].

Broader applicability: the Medpipe microservice is not limited to Pannotator; it can also be integrated with other genomic analysis tools, providing flexibility for various research applications [1,2].

By integrating Pannotator with Medpipe through a robust microservice, researchers are equipped with powerful tools for comprehensive genomic analysis, paving the way for advancements in bioinformatics and related fields.

3.8. Error management in annotation and prediction

The system utilizes an integrated microservice architecture to manage errors during annotation and prediction processes. Key features include:

- Status retrieval: users can query the status of their annotation or prediction requests, allowing them to identify if an error has occurred.
- Error logging: detailed logs are maintained to provide insights into any issues that arise during processing, helping users diagnose problems effectively.
- Documentation availability: comprehensive endpoint documentation and request examples are provided, which help users understand the expected inputs and outputs and reduce the likelihood of errors.
- Integration of microservices: The Medpipe microservice enhances the Pannotator tool by automating the processing, which helps minimize manual errors.

- User notifications: the system is designed to notify users in case of errors through status updates, providing clarity on the processing state.
- Modular design: the microservice architecture allows for individual components to be updated or debugged without affecting the entire system, leading to efficient error resolution.

3.9 User notification mechanisms

The system ensures that users are notified of errors effectively through several mechanisms:

- Status updates: users receive real-time updates on the progress of their requests, which include notifications of any errors encountered during processing.
- Feedback mechanism: users can report persistent issues or errors, contributing to future enhancements in the system's error-handling capabilities.
- Community support: platforms may offer forums or community support where users can share experiences and solutions regarding common errors.

These features collectively ensure that users are kept informed and can effectively manage any issues that arise during the annotation and prediction processes.

4. Results

Medμσ provides detailed information on mature epitope density and the protein's classification in relation to its component in Gene Ontology. To depict the annotation improvements provided by Medμσ in Pannotator, we selected a region from one of our in-house genomes annotated using Pannotator. The genome region was chosen because it illustrates several situations where one can visually perceive the advantages of our visualization schema proportioned by our tool Pannotator and the addends amended by the Medμσ. The illustration in Figures 3 and 4 encompasses several membrane integral proteins, one protein potentially exposed at the bacterial membrane surface, another pool of cytoplasmic proteins, and a single secreted protein. Moreover, three proteins possess low amino acid identity compared to the proteins of the reference genome to the species, eleven proteins with more than 95% amino acid identity and size compared to the same proteins in the reference genome, and two standing between 70 and 94% of identity and size compared to the orthologues present in the reference genome. However, all this information is absent in Figure 3, the standard genome representation for GenBank files.

The only information we provided here that one could infer by analyzing Figure 3 concerns two genes with a below-average GC content region (the third and fourth genes in the forward strand). These two genes can be considered atypical to the reference genome, and one can infer low protein identity to the reference's proteins.

Figure 3. Result Only with Pannotator annotations. The standard genome representation lacks lots of feature annotations from users, making it difficult to investigate probable proteins' roles.

Figure 4. The Medμσ annotations are included in the Pannotator. The Pannotator color code: green (highly identical to a protein from the reference genome), yellow (most likely to the reference genome), and red (most unlikely to reference genome). The red ones correspond to the below-mean GC content. Signal peptides and transmembrane domains are represented in the DNA strands corresponding to the proteins in red and magenta, respectively.

The features we informed are coded as a color scheme in Figure 4. The coding sequences possessing more than 95% identity and size compared to their best Blast matches in the

reference genome are green-colored, the yellow-colored proteins are between 70 and 94% identity and size to the references' proteins, and the red-colored proteins are below 70%. The 70% is the default cut-off value adopted by our Pannotator tool to apply this color code to all annotated proteins. We should note that our default value of 70% is appropriated to unmask genes possessing unconventional GC content since the genes colored in red fall within the region of below-average GC content. This coloring schema is used by the Pannotator tool as a default. If a user does not want to keep this color code (for NCBI submission purposes), a simple color tag removal drops the color code from a genome annotated by Pannotator.

Besides an efficient color code for coding sequences, the Pannotator can now incorporate the local subcellar predicted to all coding sequences by representing hydrophobic regions indicated by the Medμσ. The Pannotator stores this data projection in the DNA strand corresponding to the coding sequence, using the GenBank features called 'misc_feature' or miscellaneous feature. Hydrophobic regions representing signal peptides, a secretion signal, are red-colored. However, this color does not conflict with red-coding sequences since one data is written in the DNA strand and the other in the reading frames. The hydrophobic regions occurring in batches through DNA strands are a signal of membrane or surface-exposed proteins. One can differentiate both according to the number and location of the hydrophobic regions. For instance, in Figure 4, the six coding sequences located at the forward strand have hydrophobic domains along all its extensions, a signal of membrane integral proteins. However, the second protein from the right to the left has no hydrophobic regions covering all the protein extensions, allowing for the possibility of a surface-exposed protein. The remainder of proteins in Figure 4 can be cytoplasmic since no hydrophobic domains have been predicted along the entire protein extensions.

Other data incorporated into the Pannotator results by Medμσ is a statistic depicting proteins more prone to success in crafting a vaccine or diagnostic test for pathogenic bacteria. This statistic is called Mature Epitope Density (MED). An article by the principal investigator of this work first defined the MED [5]. The idea is that secreted and surface-exposed proteins are more prone to strongly containing binder epitopes to the MHC molecules. As many epitopes are predicted in a protein, the odds of starting a host's immunological response to infections are greater. The coding sequences annotated by the Pannotator and predicted as secreted or surface exposed by Medpipe receive a GenBank feature note indicating the MED statistic. This data is normalized by the greater MED value obtained among all secreted and surface-exposed proteins, so the better targets for a vaccine have the MED close to the value of one. To consult the MED using Artemis, users can ask the software to show the properties of coding sequences containing a few hydrophobic motifs (transmembrane and signal peptide motifs) at the beginning of a coding sequence. The Artemis user will find in the note feature a text like this: "Mature Epitope Density (MED): 1.0" or other values. To avoid a blind search strategy, an Artemis user can also ask the software to list all proteins contained inside the note feature with the keyword "MED".

One should pay attention to the previous description of "a few hydrophobic motifs" that we used in the last sentence. The reason is that a coding sequencing with many such motifs is probably not secreted or surface-exposed but is membrane integral. All membrane proteins

have larger MED statistics, even greater than secreted and surface-exposed proteins. However, if we are trying to produce a vaccine using membrane proteins, this task promises to be harder, considering wet lab techniques to isolate and express these proteins. Because of the enormous obstacles to using membrane proteins as vaccine targets, we opt not to show MED for membrane proteins. We also choose not to include cytoplasmatic proteins due to low MED and being most of a predicted proteome, which poses extra and unjustified load to our hardware executing the MED predictions.

5. Discussion

The Medpipe integration with Pannotator enriched our genomic analysis, providing additional information that is essential for understanding the biology and functionality of the organisms under study. This data has the potential to drive future scientific discoveries and research, contributing to a more comprehensive and detailed understanding of genomics and proteins of interest.

The Pannotator allows for rapid and correct genome annotation since one uses the accepted reference genome from the NCBI as the only source of annotation to a novel genome. Moreover, after a genome deposit, a research team will prefer to use the Pannotator copy of a genome since it contains several data that are absent at the NCBI site for that genome. Knowledge about proteins' subcellular location allows for formulations of hypotheses about the protein's relationship, therefore impacting the fields of genomics and bioinformatics.

The main limitation of this study is the lack of a predictor for non-classical secreted proteins. However, to continue this work, we intend to incorporate an in-house predictor for this purpose into Medpipe in our next software release.

The fact that the literature review on Medpipe-like tools in the last ten years has returned so few results makes us endorse the hypothesis that pharmaceutical companies would not be interested in implementing vaccines against many of the infectious diseases transmitted by bacteria for which we have today's antibiotics. The development of a vaccine can take decades and involve billions of dollars in spending. The reason for this lack of interest in solutions against infectious diseases would be in the sale of antibiotics and anti-inflammatories, one of the main pillars of support for the pharmaceutical industries, and the employability of physicians who, in theory, control the release of these drugs. The sale of antibiotics never stops, while a few doses of a vaccine mean the end of the trade of millions or billions in antibiotics. Another probable reason is the fact that many bacterial diseases impact underdeveloped or developing countries (diseases neglected by rich countries). The main stakeholders in vaccines are countries that cannot create them. Still, these countries will be long-standing customers of pharmaceutical companies by buying antibiotics and anti-inflammatories. The reason is that many bacterial diseases impact underdeveloped or developing countries (diseases neglected by rich countries). The main stakeholders in vaccines are countries that cannot create them. Still, these countries will be long-standing customers of pharmaceutical companies by buying antibiotics and anti-inflammatories.

Notably, even after eleven years of publication, the Medpipe concept is still innovative, and there are no other solutions that counter it. Therefore, the Medμσ is not a new algorithm but another pathway to access the Medpipe software without the limitations of third-party authorial rights.

5.1. Limitations of the current system

Lack of comprehensive tools: the literature review indicates a scarcity of available software tools like Medpipe for identifying candidate proteins for vaccines or diagnostic targets against pathogenic bacteria [5].

Non-Open-source nature: Medpipe is non-free software, limiting access for researchers who prefer open-source alternatives and thereby restricting its usability in certain research environments [2].

Geographical disparities: many bacterial diseases primarily affect underdeveloped or developing countries, which are often neglected by wealthier nations where research funding and interest are concentrated [13].

Innovation stagnation: despite the innovative concept of Medpipe, there have not been significant advancements or new algorithms developed in the field over the past eleven years, indicating a stagnation in research progress [13].

Limited annotation capabilities: while Medpipe offers some functional annotation, further integration of pathogenicity annotations that can enhance the understanding of protein functionalities remains necessary [7].

Microservices implementation challenges: the implementation of microservices for biological analysis is still evolving, and achieving seamless operation among different services can pose technical challenges [1].

5.2. Potential areas for future research

Development of open-source alternatives: research could focus on creating open-source tools that replicate the functionality of proprietary software like Medpipe, making them more accessible to the scientific community.

Vaccine development incentives: investigating new business models that align pharmaceutical interests with public health needs could promote the development of vaccines for neglected diseases [13].

Enhanced data integration tools: future research could aim to create more sophisticated integration tools that allow seamless interoperability between different genomic and proteomic analysis platforms [2].

Focus on neglected diseases: targeting research efforts towards bacterial diseases that disproportionately affect underdeveloped regions could lead to innovative solutions that address global health disparities [13].

Annotation and pathogenicity studies: expanding the scope of functional annotation to include more detailed insights into protein pathogenicity may yield better vaccine candidates [7].

Collaboration between sectors: encouraging partnerships between academia, industry, and governments could foster a more collaborative approach to infectious disease research and vaccine development [13].

Longitudinal studies on antibiotic resistance: research could focus on the long-term implications of antibiotic use, exploring alternatives such as vaccines that could reduce reliance on antibiotics and mitigate resistance [13].

User-centric tool development: future tools should be designed with user feedback in mind, ensuring they meet the practical needs of researchers in various fields [2].

Funding for innovative research: advocacy for increased funding towards innovative research in underfunded areas, such as vaccine development for bacterial diseases, could lead to breakthroughs in public health [13].

6. Conclusion

The integration of Medpipe into Pannotator represents an advancement in the ability to analyze and interpret genomic data. The results obtained show the effectiveness of this integration by providing detailed data on gene annotation and genetic products, including Mature Epitope Density (MED) and protein classification in relation to its component in Gene Ontology (GO).

The automation of the process, provided by the integration with the microservice Medμσ, simplifies and streamlines genomic analyses. The applicability of the integration extends to projects and research that require a deeper, more comprehensive understanding of genomic data. The availability of Medμσ endpoints capabilities allows for future expansions and integrations with other tools and services, providing a flexible and adaptable platform to the needs of ever-evolving genomic research. Therefore, the successful integration of Medpipe into Pannotator using a microservice called Medμσ represents an advancement in capacity analysis and interpretation of genomic data, providing additional data to a more efficient genomic annotation.

Conflicts of interests

The authors declare that there are no conflicts of interest regarding the publication of this article.

Ethical statement

It is not applicable to the scope of this article.

Authors' contribution

Conceptualization, A.S.; methodology, A.S., and R.G.; software, A.S., and R.G.; validation, A.S.; formal analysis, A.S.; investigation, A.S.; resources, A.S.; data curation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S.; visualization, A.S., and R.G.; supervision, A.S.; project administration, A.S. All authors have read and agreed to the published version of the manuscript.

References

[1] Gangotia D, Gupta A, Mani I. Role of Bioinformatics in Biological Sciences. In *Advances in Bioinformatics*, 1st ed. Singapore: Springer, 2021, pp. 37–57.

[2] Zhang Y, Zhang Q, Zhou J, Zou Q. A survey on the algorithm and development of multiple sequence alignment. *Brief. Bioinform*. 2022, 23(3):1–16.

[3] Costa SS, Guimarães LC, Silva A, Soares SC, Baraúna RA. First Steps in the Analysis of Prokaryotic Pan-Genomes. *Bioinform. Biol. Insights* 2020, 14:1177932220938064.

[4] Santos A, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, *et al.* PANNOTATOR: an automated tool for annotation of pan-genomes. *Genet. Mol. Res*. 2013, 12(3):2982–2989.

[5] Santos A, Pereira V, Barbosa E, Baumbach J, Pauling J, *et al.* Mature Epitope Density—A strategy for target selection based on immunoinformatics and exported prokaryotic proteins. *BMC Genomics* 2013, 14 Suppl 6(Suppl 6):S4.

[6] Fowler M. Microservices. 2014, Available: https://martinfowler.com/articles/microservices.html (accessed on 20 March 2024).

[7] Hossain MD, Sultana T, Akhter S, Hossain MI, Thu NT, *et al.* The role of microservice approach in edge computing: Opportunities, challenges, and research directions. *ICT Express* 2023, 9(6):1162–1182.

[8] Velepucha V, Flores P. A Survey on Microservices Architecture: Principles, Patterns and Migration Challenges. *IEEE Access* 2023, 11:88339–88358.

[9] Fowler SJ. *Microsserviços prontos para a produção: Construindo sistemas padronizados em uma organização de engenharia de software*, São Paulo: Novatec Editora, 2019.

[10] Newman S. *Building Microservices*, 1st ed. Sebastopol: O'Reilly Media, 2015.

[11] Liwei G, Anush M, Li-Heng C, Vinicius C, Aditya M, *et al.* Rebuilding Netflix Video Processing Pipeline with Microservices | by Netflix Technology Blog | Netflix TechBlog. 2024, Available: https://netflixtechblog.com/rebuilding-netflix-video-processing-pipeline-with-microservices-4e5e6310e359 (accessed on 20 March 2024).

[12] H J. 4 Microservices Examples: Amazon, Netflix, Uber, and Etsy. 2023, Available: https://blog.dreamfactory.com/microservices-examples/ (accessed on 20 March 2024).

[13] Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014, 30(14):2068–2069.

[14] Tobes R, Pareja-Tobes P, Manrique M, Pareja-Tobes E, Kovach E, *et al.* Gene calling and bacterial genome annotation with BG7. *Methods Mol. Biol*. 2015, 1231:177–189.

[15] Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep*. 2015, 5:8365.

[16] Leung E, Huang A, Cadag E, Montana A, Soliman JL, *et al.* Protein Sequence Annotation Tool (PSAT): A centralized web-based meta-server for high-throughput sequence annotations. *BMC Bioinformatics* 2016, 17(1):43.

[17] Kremer FS, Eslabão MR, Dellagostin OA, da Pinto LS. Genix: A new online automated pipeline for bacterial genome annotation. *FEMS Microbiol. Lett*. 2016, 363(23):fnw263.

[18] Casimiro-Soriguer CS, Muñoz-Mérida A, Pérez-Pulido AJ. Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. *Proteomics* 2017, 17(12).

[19] Mende DR, Letunic I, Huerta-Cepas J, Li SS, Forslund K, *et al.* ProGenomes: A resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res*. 2017, 45(D1):529–534.

[20] Tanizawa Y, Fujisawa T, Arita M, Nakamura Y. Generating publication-ready prokaryotic genome annotations with DFAST. In *Methods in Molecular Biology*, United States: Humana Press, 2019, pp. 215–226.

[21] Sallet E, Gouzy J, Schiex T. EuGene: An automated integrative gene finder for eukaryotes and prokaryotes. In *Methods in Molecular Biology*, United States: Humana Press, 2019, pp. 97–120.
[22] Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, *et al.* DescribePROT: Database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.* 2021, 49(D1):298–308.
[23] De Souza EV, Dalberto PF, Machado VP, Canedo A, Saghatelian A, *et al.* µProteInS-A proteogenomics pipeline for finding novel bacterial microproteins encoded by small ORFs. *Bioinformatics* 2022, 38(9):2612–2614.
[24] Jaiswal V, Chanumolu SK, Gupta A, Chauhan RS, Rout C. Jenner-predict server: Prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics* 2013, 14(1):211.
[25] Nagpal G, Usmani SS, Raghava GP. A web resource for designing subunit vaccine against major pathogenic species of bacteria. *Front. Immunol.* 2018, 9:2280.
[26] He Y, Racz R, Sayers S, Lin Y, Todd T, *et al.* Updates on the web-based VIOLIN vaccine database and analysis system. *Nucleic Acids Res.* 2014, 42(D1):1124–1132.
[27] Abby SS, Denise R, Rocha EP. Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder Version 2. In *Methods in Molecular Biology*, United States: Humana Press, 2024, pp. 1–25.
[28] Yero D, Conchillo-Solé O, Daura X. Antigen discovery in bacterial panproteomes. In *Methods in Molecular Biology*, United States: Humana Press, 2021, pp. 43–62.
[29] Saudate A. *APIs REST: Seus serviços prontos para o mundo real*, São Paulo: Casa do Código, 2021.
[30] The Spring Team. Spring Boot Reference Documentation (Version 3.1.6). 2023, Available: https://docs.spring.io/spring-boot/docs/3.1.6/reference/html/ (accessed on 20 March 2024).
[31] Miller FP, Vandome AF, McBrewster J. *Apache Maven*, São Paulo: Alpha Press, 2010.
[32] H2 Database Engine - Cant. 2024, Available: http://www.h2database.com/html/main.html (accessed on 20 March 2024).