UNIVERSITY OF GRANADA

MAJOR IN COMPUTER SCIENCE

GeneSys

A BIOINFORMATIC TOOL FOR GENOMIC DATA ANALYSIS

Author: Bruno Otero Galadí

Supervisor: Dr. Fernando Berzal Galiano

August, 2024. Granada, Spain.

Genesys: A bioinformatic tool for genomic data analysis

Bruno Otero Galadí

Keywords: reverse transcriptase, amino acid, molecular biology, nucleotide, protein, DNA, RNA

Abstract:

Recent decades' advancements in biological research have brought numerous benefits to our understanding of nature and society's progress. However, these advancements have also generated a vast amount of biological data that must be processed quickly to remain valuable for researchers. If the required speed in processing this data is not achieved, it could become a bottleneck, potentially slowing the current rate of scientific discoveries.

The majority of the problems are related to the preprocessing of big amounts of biological data stored in public databases, which are continuously updated to locate more and more examples of genomic information coming from all kind of sources. So, if researchers without advanced programming knowledge want to dive into these databases in search for a specific kind of genomes, they must be able to manipulate the data in a way that allows them to repeat the process with as many parameters as needed. Additionally, researchers may need to process the data through a series of tasks that must be separated and executed sequentially. This is where GeneSys comes into play.

GeneSys is a modular and scalable software tool with a user-friendly interface that allows researchers to define tasks within a workflow that can be executed and redefined freely in order to satisfy their researching needs, regardless of complexity.

The GeneSys software is designed to have a basic first layer that defines how tasks and workflows are related to each other. This structure allows developers to create modules that would address specific problems. This work includes am initial module designed to solve a real life issue involving reverse transcryptases, also known as RTs, a unique kind of proteins with significant research potential, many aspects of which remain unexplored. Such proteins are currently being studied by Dr. Francisco Martínez-Abarca Pastor at La Estación Experimental del Zaidín (EEZ) in Granada, Spain. The implemented module will help Martínez-Abarca to efficiently face his investigations involving RTs.

Genesys: una herramienta bioinformática para el análisis de datos genómicos

Bruno Otero Galadí

Palabras clave: reverso transcriptasa, aminoácido, biología molecular, nucleótido, proteína, ADN, ARN

Resumen:

Muchos avances se han dado en las últimas décadas en la investigación biológica, todos ellos aportando progresos en la comprensión de la naturaleza y en el desarrollo de la sociedad. No obstante, estos avances han provocado la necesidad de procesar cada vez más datos biológicos a un ritmo que debe permanecer constante para resultar rentable. Si dicha eficiencia en el procesado de datos no se alcanza, existe el riesgo de que se convierta en un cuello de botella que, llegado el momento, reduja el ritmo con el que se han producido avances en esta materia hasta ahora.

La mayoría de los problemas van de la mano al preprocesamiento de información genética contenida en diversas bases de datos, que además se incrementa en volumen con el paso del tiempo, a medida que se descubren nuevos genomas. Cualquier persona investigadora que carezca de un nivel alto de programación y desee emplear información de una base de datos para acometer una tarea va a necesitar disponer de un mecanismo que le permita repetir el proceso aplicado a los datos tantas veces como desee, así como subdividir el trabajo a realizar en tareas distintas, en caso de que quiera separarlas en el tiempo y ejecutarlas una a una. Es aquí donde entra GeneSys.

GeneSys es una aplicación modular y escalable con una interfaz de usuario fácil de usar, enfocada en ayudar en las tareas de investigación de datos biológicos. Permite a un usuario general definir tareas dentro de un flujo que podrá ejecutar y modificar según sus necesidades.

GeneSys incorpora una capa software básica que define la forma en la que las tareas y los flujos de tareas se relacionan en la aplicación. Partiendo de ahí, es posible implementar módulos personalizados e independientes que acometan tareas según las necesidades específicas de las investigaciones que se estén llevando a cabo. Este trabajo, además de la capa básica, incluye un módulo diseñado para resolver un problema de preprocesado de datos relativo a las reverso transcriptasas, también conocidas con RTs, un tipo de proteínas con un potencial investigador enorme de las que aún no se conoce mucho. El doctor Francisco Martínez-Abarca Pastor de la Estación Experimental del Zaidín (EEZ) de Granada, España, se encarga en la actualidad de estudiar dichas proteínas. El módulo implementado le servirá para progresar en sus investigaciones.

I, **Bruno Otero Galadí**, scholar of the **computer science** university degree at the "**Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**", with a Spaniard national identification number of **75574203K**, authorize the placement of the present work at my school's library so it can be consulted by anyone who wishes to.

Signed: Bruno Otero Galadí

Granada, on September the 1st of 2024.

Mr. **Fernando Berzal Galiano**, teacher of the Computing Science and Artificial Intelligence Department of the University of Granada.

Informs:

That the present work entitled as **Genesys: A bioinformatic tool for genomic data manipulation**, has been realized under his guidance by Bruno Otero Galadí, and authorizes the defense of the aforementioned work under the collegiate tribunal that might correspond.

And so that it is stated, he issues and signs the present invoice in Granada on <month> the <day> of 2024.

Supervisor:

Fernando Berzal Galiano

Acknowledgements

This work would have never existed without my supervisor, Fernando, whose suggestion to focus on bioinformatics was crucial in shaping the direction of my research. Additionally, I would not have been able to discover the significance of reverse transcryptases (RTs) and the reasons for their study without the assistance of Francisco Martínez-Abarca Pastor, a former researcher at the Estación Experimental del Zaidín (EEZ) in Granada, Spain. Francisco asked me to help him facing the RTs issue involving the preprocessing of amino acid data from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) online database. His role as a client in this work is the very reason it came into existence.

Furthermore, I want to give sincere thanks to Antonio Quesada Ramos, my former high school biology teacher, who facilitated my connection with Francisco. He is also the reason why I am so interested in biology as a field of research.

Finally, all of the time and resources I have dedicated to this matter are direct merit of my family —Dulcinea, David and Leonardo— whose support and understanding provided me with all the space I needed in order to achieve the main goals of this work. Their encouragement has been indispensable. So, thank you.

MAIN INDEX

IMAGE INDEX

TABLE INDEX

CODING INDEX

1. INTRODUCTION

1.1. A brief overview of molecular biology.

If we assume selecting breeding as a form of molecular biology researching, we can affirm that genetics has been taking part in humanity's history since, at least, the Neolithic period. But it was not until nineteenth century and the appearance of Gregor Johann Mendel's works that a first theoretical basis for the principles of heredity was set. Since then, molecular biology has become one of the most developed researching fields, being continually adapted to answer new questions and to face new challenges. As a result, molecular biology went from studies about peas to relatively recent works that suggest the existence of life beyond planet Earth, always being strongly correlated to chemistry and incorporating key discoveries like the DNA structure, which also paved the way for other numerous applications such as the Polymerase Chain Reaction (PCR) with a major relevance in the understanding and diagnosing of several diseases, or the Human Genome Project in 1990[4]. All of this improvements have provided invaluable benefits for society. Biology and specially molecular biology are not just fields with history. They are fields with future.

Before we get deeper into biological concepts, we should remember what DNA, RNA and proteins are and how they are related to each other. DNA[5] and RNA[6] sequences are identified as sequences made up of repetitions of up to four nucleotides, represented as A, C, T, G for DNA strings and A, C, U, G for RNA ones. Apart from the composition, the main differences between both structures involve their spacial distributions (with a double-helix polymer structure for DNA and a single-stranded biopolymer for RNA) and their biological functions, with DNA serving as a codification of the genetic information and RNA using DNA to synthesize the proteins that are stored in cells[7]. Proteins are chemical components made of elements called amino acids. The amino acids that might be found in proteins' cells differ between species, but there are no more than twenty different amino acids that occur naturally in any living being's proteins[8].

To sum up, DNA defines the genetic composition of a living being, and RNA replicates that composition in order to define the structure of proteins. But, what is the mechanism that translates RNA into proteins? The nucleotides contained in a RNA sequence (and therefore in its equivalent DNA sequence) are read in intervals of three each, and they can be read starting from anyone of the first three nucleotides that compound the aforementioned RNA (or DNA) sequence, onward and backward, which gives us up to six different ways of getting a protein from a same RNA string. A protein is properly identified when, using one of those lecture ways, a specific set of three nucleotides that marks the end of the lecture is found. The available combinations that correspond to that case are: "UAA", "UAG" and "UGA" ("TAA", "TAG" and "TGA" for DNA sequences). This sets of nucleotides are called stop codons, and correspond to a certain amino acid that serves as a delimiter of the protein[9].

1.2. Reverse transcryptases.

Reverse transcription refers to the process of turning specific RNA sequences into DNA. Not all RNA strings are valid for this issue, and those that indeed are are formally called "RNA-dependent DNA polymerases", "reverse transcryptases" or "Rts"[10]. As not all RNA strings serve as RTs, they need to be specifically recognized before researchers start experimenting with them. As we have stated before, RNA sequences are translated into proteins. That process can be done backwards, too, which means that it is possible to find a specific protein that is configured by a RNA sequence that in fact is a RT. In other words, we

can identify RTs by observing protein structures.

Reverse transcryptases have remarkable biotechnological applications, such as molecular cloning strategies or in the field of synthetic biology. But the most important use they have provided to humanity, or at least the most widespread one, might be the detection of viral RNA in SARS-CoV-2 testings[1], as they serve as a key element in the propagation of genetic elements across specific DNA structures.

Since 2020 COVID-19 pandemic, the lock-down and the infection waves, the interest in molecular biology seems to have gained so much popularity, being mentioned in the news, in social networks or even at the dinning room with our families. However, and despite the crucial role they have played throughout all these years, reverse transcryptases have not become that popular. And as researchers and diverse studies point out that pandemics would be more common in the future, it is quite clear that RTs will keep being at the spotlight of scientific investigations. The main arguments that are exposed to support the assumption of pandemics becoming more likely to happen concern topics such as climate change[2], the destruction of the environment or the increasing contact between humans and disease-harboring animals[3].

In a post-COVID world, it is crucial to be prepared for upcoming similar events. RTs take part in that process by playing a key function in PCRs, which play a potentially high disease detection role.

1.3. When more is less. Current problems involving genomic databases.

Nowadays, biologists tend to work obtaining their genetic data from enormous public domain databases whose volume of biological information is increasing at a relatively faster rhythm than the stored datasets of other scientific disciplines, with the amount of raw data corresponding to genome sequencing experiencing the biggest growth along with exome sequencing data[11]. This has led to an overwhelming amount of genomic data that needs to be correctly preprocessed in order to start searching for valuable knowledge. And RTs are taking part in that problem, too, as new examples of them are being included in those databases month by month, increasing the difficulty to distinguish which are recent discoveries from those which are not, as well as requiring more complexity in the computing of all the existing samples in order to identify common patterns between them.

Francisco Martinez-Abarca Pastor is a researcher from the Estación Experimental del Zaidín (EEZ) in Granada, Spain. Among his current issues there is the exploration of RTs' datasets in search of undiscovered correlations between reverse transcriptase samples that are separated in evolutive terms. In the year 2019, he supervised a work involving RTs' written by the former postgraduate degree student Mario Rodríguez Mestre. The aforementioned work was entitled "Analysis of Novel and unexplored groups of prokaryotic Reverse Transcriptases" and consisted of the extraction of all the available datasets of RT's stored in certain databases, its preprocessing, its classification through clustering algorithms and the seeking of undiscovered common behavior patterns between the RT's contained in each cluster.

The results of the study were considered successful, and Francisco decided he would repeat the experiment once the databases were updated with new samples. The problem is that in order to repeat the process, all the steps of preprocessing the raw data and applying the clustering had to be done manually again, which implied to look for software tools that could work with compatible data, apart from . All of that required so much effort in terms of time to be worth, so even though Francisco wished to repeat the study, he was not able to do so.

But if he had a tool that at least could automatize the preprocessing of the raw data downloaded from the databases just as Mario did back in time, he would be capable of make the same experiment whenever he wanted, so in the long term he could exploit all the advantages RTs have. GeneSys serves as a software tool that solves Francisco's issue.

1.4. Memory structure.

In order to facilitate the reader to navigate throughout this work, here there are mentioned the main parts of it and what does each expose:

- **Introduction**: this chapter exposes briefly the context in which GeneSys is made. It provides a generic view of the problem as and helps those readers with basic biology knowledge to understand what reverse transcriptases are and what they are used for.
- **Objectives**: includes the objectives to accomplish with GeneSys development. Such accomplishments will be analyzed in the "Conclusions and future work" section.
- **Planning**: organization, task to do, estimated developing time and hypothetical budget it might require.
- **Problem analysis**: statement of the preprocessing tasks that GeneSys must accomplish and the biological reasons that justify why they must be done that way.
- **Architecture and design**: it provides various diagrams that explain the architecture of the application. Also, justifies why that architecture has been chosen and what requirements are crucial to satisfy.
- **Implementation**: abstract of GeneSys' coding process and the development stages that occurred while implementing the tool.
- **GeneSys user's guide**: A friendly-user manual that explains how to use the app. It is aimed to be understood by any researcher how wishes to employ GeneSys in their investigations.
- **Conclusions and future work**: it compares the accomplished objectives against those exposed in the "Objectives" section. It also proposes improvements that can be made to the application in the future.

2. OBJECTIVES

The main objective is to provide a functional application that can be intuitively employed by an user experienced with biological terminology, but with a basic programming knowledge, which accomplishes properly the issue of RTs' preprocessing. In order to achieve the proposed goal, we can distinguish the following objectives in the development of the application:

1. To develop a bug free logic that works exactly as the user needs it to work.
2. To provide an intuitive and attractive friendly-user interface.
3. To properly recognize which tasks must be automated by the application, so that the user receives a window to modify certain parts of the preprocessing of the RTs in order to adapt their experiments freely while not losing the automatization benefits that GeneSys provides.
4. To provide a software framework that unifies all the preprocessing tasks in an unique execution context. In other words, we do not want the user to open any other application than GeneSys to achieve the proposed preprocessing tasks.
5. To make an application that does not freeze or crash when it is executing a long task.
6. To give the user the capacity to apply changes to the task that it is going to be executed, such as the pathnames where to save the results. We assume that it would be always better to give the user as much freedom as possible when defining the parameters of the preprocessing tasks.
7. To properly inform users about what is happening in the preprocessing of the RTs, so that they can study the results returned at all the steps of the process and draw their own conclusions for their research.

3. PLANNING

4. PROBLEM ANALYSIS

Parte de análisis: en vez de posibles casos de uso, límitate a analizar el problema a resolver. Por qué una aplicación de escritorio. Qué implicaciones tiene que sea de escritorio?

amino acid sequences that work as baits for RTs (in other words, a very large string of amino acid bases that potentially contains RTs within them and also stores a known protein that is employed to recognize the aforementioned long amino acid sequence as a whole)

5. ARCHITECTURE AND DESIGN

Pasar a resaltar la necesidad de implementar una aplicación escalable que resuelva más de un problema, porque la biología no se detiene y sería contraproducente y poco eficiente no hacer algo más genérico. En la parte de diseño ya sí puedes explicar que lo has organizado con vistas a poder incluir nuevos módulos en el futuro.

6. IMPLEMENTATION

7. GENESYS USER’S GUIDE

8. CONCLUSIONS AND FUTURE WORK

9. BIBLIOGRAPHY

1. **Reverse Transcriptases: From Discovery and Applications to Xenobiology**
2. **Factors that may predict next pandemic**
3. **Statistics Say Large Pandemics Are More Likely Than We Thought**
4. **Methods in molecular biology and genetics: looking to the future**
5. **DNA chemical compound**
6. **RNA chemical compound**
7. **What Is the Difference Between DNA and RNA?**
8. **Protein. Biochemistry**
9. **Translation of RNA to Protein**
10. **Reverse Transcription—A Brief Introduction**
11. **Data volume growth in genomics versus other disciplines**