UNIVERSITY OF GRANADA

MAJOR IN COMPUTER SCIENCE

GeneSys

A BIOINFORMATIC TOOL FOR GENOMIC DATA MANIPULATION

Author: Bruno Otero Galadí

Supervisor: Dr. Fernando Berzal Galiano

August, 2024. Granada, Spain.

Genesys: A bioinformatic tool for genomic data manipulation

Bruno Otero Galadí

Keywords: reverse transcriptase,

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 informática para la manipulación de datos genéticos

Bruno Otero Galadí

Palabras clave: reverso transcriptasa,

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

What are proteins?????

Reverse transcryptases (also known as “RNA-dependent DNA polymerases”, or RTs) consist of a particular kind of proteins characterized for their remarkable biotechnological applications, such as molecular cloning strategies and 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 and 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 potentially high disease detection role.

At the same time, nowadays biologists tend to work obtaining their genetic data from enormous public domain databases whose volume of biological information is continually increasing at a estimated rhythm of (look for sources), which leads to an overwhelming amount of raw data that need to be correctly preprocessed to obtain useful information in order to start searching for valuable knowledge. And RTs are taking part in that problem, too, as new amino acid sequences that work as baits for RTs (in other words, a very long string of amino acid bases that potentially has RTs within them and also stores a recognizable short protein that is employed to recognize the aforementioned long amino acid sequence as a hole) are being included in those databases, increasing the difficulty to work with them.

Francisco Martinez-Abarca Pastor, qué hace qué estudia qué hizo con Mario qué experimento hizo a qué conclusiones llegaron y por qué quiere repetirlo y qué puedo aportar yo a ese proceso

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.

2. OBJECTIVES

3. PLANNING

4. PROBLEM ANALYSIS

5. ARCHITECTURE AND DESIGN

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**