DESARROLLO DE UNA HERRAMIENTA DE SOFTWARE LIBRE PARA EL MODELADO DE LA EXPRESIÓN GÉNICA DE POBLACIONES CELULARES

BRAYAM ANDRES SAAVEDRA ARCE
2150058

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE AUTOMÁTICA Y ELECTRÓNICA
PROGRAMA INGENIERÍA BIOMÉDICA
SANTIAGO DE CALI
2021

DESARROLLO DE UNA HERRAMIENTA DE SOFTWARE LIBRE PARA EL MODELADO DE LA EXPRESIÓN GÉNICA DE POBLACIONES CELULARES

BRAYAM ANDRES SAAVEDRA ARCE

Proyecto de grado para optar al título de
Ingeniero Biomédico

Director
ANDRES MAURICIO GONZALEZ VARGAS
Doctor en Ingeniería Electrónica, Informática y Eléctrica

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE AUTOMÁTICA Y ELECTRÓNICA
PROGRAMA INGENIERÍA BIOMÉDICA
SANTIAGO DE CALI
2021

Nota de aceptación:

Aprobado por el Comité de Grado en cumplimiento de los requisitos exigidos por la Universidad Autónoma de Occidente para optar al título de Ingeniero Biomédico

LILIAN SOFÍA SEPÚLVEDA SALCEDO
Jurado

PAOLA ANDREA NEUTA ARCINIEGAS
Jurado

Santiago de Cali, 16 de junio de 2021

“El día en que dejas de aprender es el día en que comienzas a decaer.”

Isaac Asimov

AGRADECIMIENTOS

Gracias infinitas a mis padres; especialmente a mi madre, a la cual le debo todo lo que he sido, lo que soy y lo que seré. Muchas gracias también a los amigos y compañeros que conocí durante mi vida universitaria; porque de cada uno de ellos de alguna forma aprendí y compartí el conocimiento adquirido en la universidad. Por último, más no menos importante, gracias totales a cada uno de los docentes que me transmitieron sus conocimientos y sabiduría tanto a nivel profesional como personal, e igualmente reconocer la importancia de su maravillosa labor en la formación de los futuros profesionales que convertirán este país en un mejor lugar.

CONTENIDO

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

LISTA DE ANEXOS

pág.

RESUMEN

Los procesos biológicos al interior de las células, tales como la expresión génica, son el resultado de una serie de interacciones entre diferentes tipos de moléculas y reacciones que toman lugar dentro de un determinado sistema biológico. Dichos procesos pueden ser aproximados de forma práctica a través del uso de ecuaciones diferenciales. Sin embargo, en la experimentación, el producto de un proceso de expresión génica puede variar a lo largo de una población de células homogénea. Dicha población puede ser estudiada *in silico* mediante la estimación de los parámetros que caracterizan su comportamiento.

Los métodos computacionales por los cuales se simula y se modela la expresión de una población son tareas que requieren de un gran conocimiento tanto a nivel biológico como de los algoritmos utilizados para su aproximación. Por lo tanto, el objetivo de este trabajo es proponer una herramienta para el modelado de la expresión génica de poblaciones celulares. En primer lugar, se recopilaron los aspectos fundamentales tenidos en cuenta por el paradigma de la biología de sistemas, para plantear una metodología de diseño de la herramienta e implementar algunos algoritmos que permitan obtener una introducción en el área. En este trabajo, se desarrolla una interfaz de usuario mediante la cual las personas interesadas en aprender sobre sistemas biológicos en poblaciones celulares puedan comprender conceptos básicos sobre la definición del sistema, su simulación a partir de diferentes fuentes de variabilidad y su posterior modelado para estimar los parámetros que las caracterizan. La herramienta se implementa en software de uso libre, de tal manera que sea posible acceder a ella de forma abierta. Así, permitiendo una potencial modificación, mejora, o uso de los algoritmos que integran la herramienta.

Por último, se evalúa la usabilidad de la herramienta mediante la elaboración de la documentación de su uso y la creación de una experiencia educativa dirigida a estudiantes de Ingeniería Biomédica, en la cual se presenta todo el proceso de modelado de una población celular a través de un caso de estudio.

Palabras clave: Sistema biológico, expresión génica, ecuaciones diferenciales, simulación, modelando, inferencia de parámetros, población celular.

ABSTRACT

Biological processes inside cells, such as gene expression, are the outcome of a series of interactions between different kinds of molecules and reactions that take place in a biological system. These processes can be approximated in a practical way using differential equations. However, in experimentation, the product of a gene expression process might be different throughout a homogeneous cell population. *In silico*, such a population can be studied through the estimation of the parameters that characterize its behavior.

Computational methods used to simulate and model the expression of a population are tasks that require having a good knowledge of both at a biological level and algorithms for their approximation. Therefore, the aim of this work is to propose a tool for modeling the gene expression of cell populations. First, fundamental aspects of the biological system paradigm are gathered to present a tool design methodology and implement some algorithms that allow to introduce beginners to the field. In this work, a user interface is developed so it could help others interested in biological systems in cell populations to obtain insights about basic topics on biological systems definition, simulating from several variability sources and subsequent modelling to estimate parameters. Freely usable software is used for the tool implementation; thus, it is possible to get open access to it and potentially modify, enhance, or use the implemented algorithms.

Lastly, tool usability is assessed through the creation of user documentation and a learning experience aimed at students of Biomedical Engineering. The latter presents the general process of modelling a cell population describing a case study.

Keywords: Biological system, gene expression, differential equations, simulation, modelling, parameters inference, cell population.

INTRODUCCIÓN

La expresión génica es un proceso dinámico mediante el cual la información de un gen es usada para la síntesis de un producto génico. Dicho producto usualmente se expresa en forma de proteína y es el resultado de una serie de reacciones bioquímicas que se dan de forma estocástica en un determinado sistema biológico. Esta estocasticidad en la expresión génica afecta principalmente a los procesos de transcripción y traducción, y es la responsable de que los niveles de expresión de una proteína difieran entre una población de células idénticas genéticamente. Adicionalmente, otros factores que influyen en las diferencias de expresión génica de una célula a otra son: la disponibilidad de factores de transcripción, el número de ribosomas, la degradación de mRNA y la proteólisis (Duveau, 2018).

La variabilidad en la expresión génica conlleva consecuencias importantes para la función celular, siendo benéfico en ciertos contextos y perjudicial en otros. Algunas de los procesos biológicos en los que se puede observar dicha variabilidad incluyen la respuesta al estrés, el metabolismo, el ciclo celular y el envejecimiento. Este tipo de fenómenos biológicos y su inherente variabilidad pueden ser modelados y caracterizados por medio de una de las áreas de la biología computacional conocida como biología de sistemas. Según Klipp, Liebermeister, Wierling y Kowald (2016), la biología de sistemas se define como “la disciplina que estudia las propiedades sistémicas y las interacciones dinámicas en un sistema biológico; considérese una célula u organismo, de forma cualitativa y cuantitativa mediante la combinación de estudios experimentales con modelos matemáticos” (p. 3). Es decir, que la biología de sistemas pretende entender cómo la dinámica global de una población celular o un sistema biológico se origina a partir de las interacciones entre cada uno de sus componentes individuales (Liebermeister, 2012).

El enfoque tradicional en la investigación en biología se basa principalmente en el razonamiento lógico en torno a los datos experimentales obtenidos de un sistema biológico. Por ende, la biología de sistemas puede ser usada como una herramienta que permita interpretar grandes cantidades de datos experimentales y, por lo tanto, podría ser considerada como un complemento para el enfoque tradicional de la biología. El enfoque de la biología de sistemas proporciona un marco metodológico que consiste en comprender cómo los componentes de un sistema biológico se traducen a un modelo matemático que puede ser interpretado de forma mecanicista. A partir del modelado matemático de un sistema biológico se pueden demostrar diferentes hipótesis referentes a la naturaleza de un determinado sistema y encontrar propiedades particulares que puedan ser probadas de forma experimental. Al mismo tiempo, este enfoque proporciona un medio para realizar predicciones sobre un sistema biológico partiendo de un modelo matemático (Janzén, 2012).

1. PLANTEAMIENTO DEL PROBLEMA

La biología computacional parte del uso de herramientas informáticas y técnicas matemáticas para modelar y facilitar el entendimiento de los fenómenos biológicos. Aplicada a la medicina, constituye una alternativa rápida, de bajo costo y potencialmente efectiva a los ensayos experimentales en animales y en humanos, los cuales pueden presentar riesgos y conflictos éticos. A partir de organismos simples como por ejemplo las levaduras, se puede adquirir información a nivel molecular acerca del funcionamiento de los seres vivos que permita avances tales como el rastreo del origen de una enfermedad, desarrollo de nuevos medicamentos, comprensión de procesos evolutivos, regeneración de órganos, entre otros. Dentro del gran número de áreas desde las cuales se puede trabajar en biología computacional se encuentra la biología de sistemas, la cual comprende el estudio de los componentes y las interconexiones que conforman las redes de reacciones bioquímicas tales como la interacción entre proteínas, las rutas de señalización celular, el comportamiento de complejos celulares y particularmente la expresión génica. Esta última constituye el proceso mediante el cual se obtienen proteínas a partir de genes. Sin embargo, en una población de células, la expresión génica se ve afectada por la heterogeneidad de cada célula individual; lo cual conlleva a que la expresión de los productos finales difiera entre una célula y otra ante un mismo estímulo.

La implementación de los algoritmos que permitan la computación de procesos celulares varía ampliamente en su complejidad, la cual puede llegar a requerir de una cantidad de tiempo y trabajo considerable. Por lo tanto, se necesitan herramientas, preferiblemente en software libre, que permitan a los estudiantes enfocarse más en la naturaleza de los modelos que en la implementación de estos. De igual manera, el desarrollo de dicha herramienta en software libre fomenta su reproducibilidad y garantiza su disponibilidad a largo plazo, lo cual evita que se encuentre atada a exclusividades de fabricantes sobre productos concretos y aumente su difusión como herramienta computacional.

De acuerdo con lo anterior, se plantea como principal interrogante: ¿Cómo desarrollar una herramienta informática para el modelado de la expresión génica de poblaciones celulares, que sirva como una herramienta de apoyo en el proceso de aprendizaje de los modelos matemáticos utilizados?

2. JUSTIFICACIÓN

El modelado de procesos biológicos permite describir fenómenos biológicos exhibidos en una población celular, y observar su respuesta ante un determinado estímulo y la variabilidad de la respuesta de los individuos que componen dicha población. En el caso de la farmacoterapia, esto es de importancia cuando se necesita describir cuantitativamente las interacciones entre las enfermedades, las drogas y los pacientes; con el fin de reducir la posible resistencia a las drogas y aumentar la efectividad de un tratamiento sobre un paciente (Lavielle, 2014).

A pesar de que existen diversas herramientas de acceso libre, usadas en la biología computacional, para el modelado de procesos o sistemas biológicos; estas se encuentran enfocadas principalmente a la implementación de los algoritmos para la simulación de los modelos, por lo tanto, el uso de dichas herramientas se limita a aquellas personas que posean un amplio conocimiento tanto en el uso de la herramienta como en la complejidad de un determinado modelo.

El desarrollo de esta herramienta pretende proveer una perspectiva afable para la descripción, la implementación y el uso práctico de un modelo. De este modo, la herramienta será útil para el entrenamiento y la enseñanza de estudiantes e investigadores que se estén adentrando en el mundo de la biología computacional; puesto que les facilitara el aprendizaje del área enfocándose primordialmente en comprender la naturaleza de los modelos que se aplican para modelar un determinado sistema biológico, los parámetros que lo caracterizan y las fuentes de variabilidad que afectan su respuesta.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Desarrollar una herramienta informática, en software libre, que permita modelar procesos de expresión génica a nivel unicelular y multicelular, con el fin de ser aplicada en investigación y educación en las áreas de bioinformática y biología computacional.

3.2 OBJETIVOS ESPECÍFICOS

- Realizar un análisis de requerimientos para el software planteado, basándose en las necesidades de los posibles usuarios del sistema.

- Desarrollar el modelo conceptual de la herramienta, teniendo en cuenta los requerimientos encontrados, así como las ventajas y restricciones dadas por los modelos matemáticos actualmente utilizados para simular los procesos biológicos relevantes para el proyecto.

- Implementar el modelo de herramienta desarrollado utilizando un lenguaje de programación de uso libre.

- Validar la herramienta como parte del proceso de aprendizaje de sistemas biológicos.

3.3 METODOLOGÍA

Para el desarrollo de la herramienta se decidió trabajar con una metodología tipo cascada, la cual es usada tradicionalmente para el desarrollo de software o herramientas informáticas. Por lo tanto, se llevaron a cabo las siguientes etapas que se adaptan a los objetivos propuestos para el proyecto.

3.3.1 Etapa 1: Análisis y definición de requerimientos

3.3.1.1 Identificación de las necesidades

A partir de la descripción de la problemática se identificaron las principales necesidades de los usuarios en función de los conceptos básicos requeridos para comprender la dinámica del proceso de modelar procesos biológicos como la expresión génica en poblaciones celulares

3.3.1.2 Establecimiento de especificaciones

Una vez identificadas las necesidades, se definieron los requisitos; tal que la problemática se dividió en pequeñas unidades de solución y que fueron representadas como funciones y características que ofreció la herramienta.

3.3.2 Etapa 2: Diseño de la herramienta

3.3.2.1 Descripción de la estructura global

En esta etapa se definió la arquitectura de la herramienta en base a los requerimientos definidos en la etapa anterior. La arquitectura de la herramienta se diseñó centrándose en componentes concretos; por lo tanto, se obtuvo un diagrama de funcionamiento general para obtener la respuesta de un determinado modelo.

3.3.2.2 Definición de las unidades

Se desglosó la estructura global de la herramienta en subsistemas o unidades; de tal forma que, por separado comprendan una función específica de la herramienta y que en conjunto comprendan la función general de la misma.

3.3.3 Etapa 3: Implementación de la herramienta

3.3.3.1 Implementación de las unidades

En esta etapa se ejecutó la arquitectura de la herramienta concebida en la etapa anterior. Se llevó a cabo la programación de cada una de las unidades en el

correspondiente lenguaje de programación. Para lo cual, se realizó una explicación detallada de los principios matemáticos tenidos en cuenta a la hora de implementar las unidades que integran la herramienta.

3.3.3.2 Verificación de las unidades

Incluyó la búsqueda de errores y se ejecutaron pruebas unitarias a partir de un caso de estudio propuesto, con el cual se comprobó que las unidades estuvieran listas para ser integradas a la estructura global de la herramienta.

3.3.3.3 Integración de las unidades de la herramienta

Esta etapa comprendió la agrupación de las unidades en una interfaz gráfica de usuario (GUI), la cual fue diseñada en base a la arquitectura propuesta para la herramienta, de tal forma que permitiera comprender el funcionamiento global de la misma.

3.3.4 Etapa 4: Validación de la herramienta

3.3.4.1 Documentación técnica

Se realizó la documentación necesaria sobre el uso de la herramienta, definiendo detalladamente cada uno de los componentes integrados en la herramienta y sus características técnicas.

3.3.4.2 Pruebas de usabilidad

Se validó el funcionamiento de la herramienta mediante la creación de una experiencia educativa en la cual se presentó a un grupo de estudiantes de Ingeniería Biomédica una introducción al modelado de sistemas biológicos. Adicionalmente, se realizó un taller práctico y un quiz con el fin de evaluar la herramienta como parte del proceso de aprendizaje sobre el tema.

4. MARCO CONCEPTUAL

4.1 ANTECEDENTES

Las técnicas de experimentación modernas permiten monitorear la dinámica de la respuesta de la expresión génica a estímulos ambientales en células individuales. Las mediciones de expresión génica han revelado que, incluso en poblaciones celulares isogénicas, la dinámica de la expresión es muy variable. En los últimos años se ha dedicado un gran esfuerzo al modelo matemático de diferentes procesos celulares, entre los cuales, para la expresión génica se ha centrado en la investigación cuantitativa del origen y las consecuencias de su variabilidad. Por lo tanto, se han propuesto varios enfoques para describir las diferentes fuentes de variabilidad, y se ha puesto especial atención en encontrar el enfoque de modelamiento apropiado que se ajuste a los datos disponibles de expresión génica en una población (Cinquemani, 2019).

Para modelar la expresión génica de poblaciones celulares en (Llamosi, 2016) propusieron un nuevo método basado en el uso de modelos de efectos mixtos. El método propone un sistema para modelar cuantitativamente un modelo estándar de expresión génica en una población de células de levadura mediante la inferencia de distribuciones multidimensionales que describen la población, y la derivación de parámetros específicos para células individuales. La aplicación del modelo se realiza a un subsistema del proceso de osmorregulación en levadura, el cual es iniciado por la ruta de señalización HOG como respuesta ante un choque de estrés hiperosmótico. Durante este proceso, se induce la expresión del gen STL1, el cual codifica una proteína simportadora de glicerol en la membrana plasmática, sin embargo, en el experimento fue modificado para codificar yECitrine, una proteína fluorescente que permite monitorear la respuesta regulatoria de cada célula mediante la medición de sus niveles de fluorescencia. Como resultado, demuestran que existe una relación directa entre la individualidad inferida en la expresión con características medibles del fenotipo y fisiología celular (ej.: dinámica de la expresión génica, tamaño celular, tasa de crecimiento, relación madre-hija). Por lo tanto, atribuyen los parámetros del modelo, relevantes biológicamente, a características específicas para células individuales. Una extensión del método descrito anteriormente se llevó a cabo por (Marguet, 2019). En este caso, se estudiaron las fuentes de variabilidad que afectan la herencia de los factores de expresión génica durante la división celular. En dicho estudio parten de un modelo de transcripción y traducción de la expresión génica, para proponer un modelo estocástico de la evolución de la dinámica de la expresión génica en una población de células divisorias. Basado en dicho modelo, desarrollaron un método para la cuantificación directa de la herencia y la variabilidad de los parámetros cinéticos de expresión génica a partir de datos de expresión génica unicelulares y de linaje. El

método desarrollado extiende el enfoque del modelo de efectos mixtos introduciendo un modelo que relaciona explícitamente los parámetros de las células madre con los de las células hijas en términos de un proceso autorregresivo (AR). De este modo, formula una estimación de la herencia y variabilidad en la división a través de la identificación de los parámetros del proceso AR. Los resultados obtenidos demostraron que mediante este método se pueden obtener estimaciones imparciales de los parámetros de herencia madre-hija. Adicionalmente, muestran que ante un determinado estimulo, los parámetros de las células hijas se encuentran determinados en gran medida por los de la madre. Esto confiere relevancia a su método en cuanto a la correcta evaluación del inicio de la variabilidad de la expresión génica y el estudio de la transmisión de factores reguladores.

En trabajos recientes se ha demostrado la factibilidad experimental de controlar la expresión génica en tiempo real mediante el uso de controladores; de tal forma que se tome el control de la expresión de proteínas a nivel intracelular permitiendo probar la flexibilidad de la dinámica unicelular. En (Maruthi, 2014) idean un sistema de control estocástico basado en un modelo de control predictivo. El sistema parte de un modelamiento estocástico de la dinámica de respuesta génica, y lo combina con un controlador de retroalimentación de estado y un método de estimación en tiempo real para determinar y compensar las variables de estado no observadas. Aplicando dicho sistema a modelos de expresión génica inducible por estrés hiperosmótico en levaduras; mostraron *in silico* el potencial de su aproximación. Para ello, en tiempo real, observaron el nivel de proteína y modularon la inducción génica en base al nivel de expresión génica deseado. De este modo, lograron mantener el nivel promedio de expresión de una proteína de fluorescencia en un valor objetivo durante varios periodos de tiempo e incluso lograron seguir perfiles de fluorescencia que varían en el tiempo con buena precisión cuantitativa.

Existen diversas herramientas que pueden ser enfocadas en la implementación de modelos de sistemas biológicos. Entre estas se encuentran las herramientas *MLXPLORE, Simulx,* y *MONOLIX*. Estas herramientas han sido desarrolladas por *LIXOFT* e *INRIA* en lenguaje *MLXTRAN* específicamente para implementar modelos de efectos mixtos. *MLXPLORE* es una herramienta que permite visualizar tanto el modelo estructural como el modelo estadístico, lo cual es de gran importancia al momento de implementar el enfoque poblacional de un modelo. De igual manera, permite visualizar el impacto de las covariables y la variabilidad en la estimación de los parámetros de un modelo. *Simulx* permite simular modelos complejos para datos longitudinales. Por su parte, *MONOLIX* es una GUI de modelamiento para modelos de efectos mixtos. Esta plataforma permite realizar diferentes tareas, tales como: estimación de parámetros poblacionales e individuales, evaluación de modelos mediante gráficos de diagnóstico y criterios de selección, y simulación de datos nuevos (Lavielle, 2014).

Con los avances en el modelamiento de la expresión génica, se ha incrementado la necesidad de herramientas que provean aproximaciones precisas de la dinámica estocástica de redes de regulación génica (GRNs) con un esfuerzo computacional reducido. En (Pájaro, 2017) se desarrolló *SELANSI,* una herramienta computacional para la simulación de redes regulatorias génicas multidimensionales estocásticas. *SELANSI* aprovecha las propiedades estructurales intrínsecas de las GRNs para aproximar con precisión la ecuación química maestra correspondiente mediante un modelo generalizado basado en ecuaciones integro-diferenciales parciales, el cual es resuelto mediante un método numérico semi-lagrangiano con alta eficiencia. Las redes modeladas por esta herramienta pueden involucrar múltiples genes partiendo de una descripción generalizada de la GRN, en la cual cada proteína puede interactuar con su gen correspondiente o con otros genes para regular su propia expresión. En general, esta herramienta ofrece flexibilidad total en cuanto a la topología de la red, su cinética y su parametrización, así como opciones de simulación.

4.2 MARCO TEORICO

4.2.1 Procesos celulares bioquímicos

Todos los procesos biológicos que suceden en los seres vivos se encuentran basados en la bioquímica. Por lo tanto, los sistemas biológicos se encuentran constituidos a partir de la configuración de determinadas moléculas y complejos de macromoléculas, los cuales presentan un comportamiento dinámico de naturaleza estocástica en sus interacciones, movimientos y transformaciones (Liebermeister, 2012). Un sistema biológico puede ser considerado cualquier organismo, población de células, o conjunto de biomoléculas que se encuentran altamente organizados en estructura y función, y en el cual se dan una serie de reacciones bioquímicas continuas necesarias para preservar la vida. El flujo de información de dichas reacciones se basa en el dogma central de la biología molecular, el cual de forma puntual establece cómo los genes codifican mRNA, el mRNA sirve como modelo de las proteínas, y las proteínas cumplen una función a nivel celular (Klipp, 2016). Bajo este dogma funcionan las diferentes redes de reacciones bioquímicas, de tal modo que forman sistemas de retroalimentación complejos que permiten que las células se adapten a diferentes estímulos provenientes del entorno exterior (Liebermeister, 2012). Las redes bioquímicas comprenden desde reacciones enzimáticas aisladas hasta sistemas completos de varias reacciones. Existen diversos tipos de redes bioquímicas, entre las más comunes podemos encontrar redes que comprenden procesos metabólicos, regulatorios, o de señalización. Las redes metabólicas son aquellas que degradan o sintetizan ciertos componentes esenciales dentro de un sistema biológico. Por su parte, las redes regulatorias se encargan de controlar la forma en que los genes son expresados como RNA o proteínas. Finalmente, las

redes de señalización se encargan de transmitir señales bioquímicas dentro de una misma célula o entre células diferentes.

Particularmente, las redes de regulación génica parten de proteínas, tales como los factores de transcripción, para inducir la expresión de genes que cumplen un papel importante en la transcripción de mRNA. Este último es traducido en otras proteínas que actúan como enzimas en rutas metabólicas o de señalización (González, 2014). En la Figura 1 se puede observar el proceso de expresión génica, el cual comienza con la interacción entre un promotor y una secuencia genética, para resultar en la producción de una determinada proteína (Cinquemani, 2019). De este modo, conocer el tipo de reacciones que ocurren en el interior de una célula es de gran importancia para avanzar en el proceso de modelarla y comprender los parámetros que caracterizan y predicen su comportamiento.

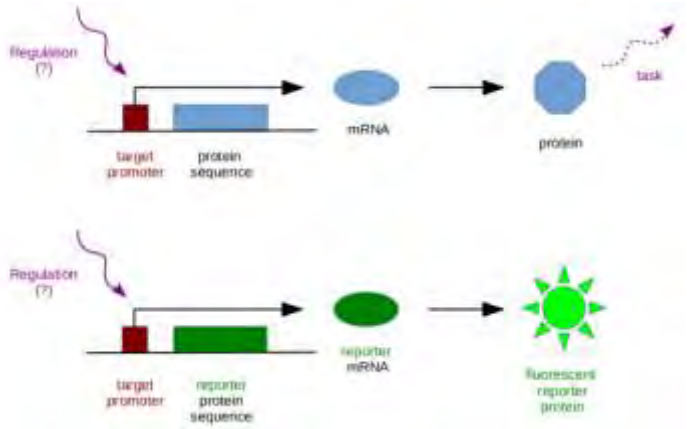

***Figura 1*. Expresión génica de una proteína.**

De "Identification, Estimation and Control of Gene Expression and Metabolic Network Dynamics" (p. 17), por E. Cinquemani, Université Grenoble-Alpes, 2019.

4.2.2 Fuentes de variabilidad en los procesos celulares

La variabilidad celular hace referencia a la predisposición que tiene una población de células idénticas genéticamente a responder de forma diferente ante un mismo estímulo o a las condiciones de un entorno extracelular. Por consiguiente, en dicha población se manifiesta una variación considerable, de una célula a otra, en los niveles de expresión de una determinada proteína o mRNA. Las principales fuentes de variabilidad en los procesos biológicos pueden considerarse de origen extrínseco o intrínseco. De manera específica, la variabilidad extrínseca se considera aquella fuente de variabilidad que surge a partir de diferencias fenotípicas de una célula a

otra, como variaciones en el tamaño celular o en las etapas del ciclo celular. Otro factor que influye en este tipo de variabilidad son las condiciones del microambiente celular relacionados con las condiciones de crecimiento de una población celular, y que genera cambios adaptativos en las células (González, 2014). Por ejemplo, la expresión de una determinada proteína se ve afectada por el número de ribosomas y el número de RNA polimerasas, las cuales varían en función del tiempo y de una célula a otra. Por su parte, las fuentes de variabilidad intrínseca surgen como consecuencia de la inherente naturaleza estocástica de las reacciones bioquímicas asociadas a la transcripción, traducción, regularización y degradación de mRNA y proteínas (Singh, 2013) (Toni, 2013). Aparte de estas fuentes de variabilidad también se considera como fuente y se tiene en cuenta, el ruido producto de los instrumentos de medición usados en la recolección de datos experimentales.

A partir del estudio de la variabilidad es posible demostrar que la dinámica de una población celular o sistema biológico es de naturaleza estocástica (Janzén, 2012). Además, el hecho de que las moléculas involucradas en los procesos biológicos, como el mRNA, se encuentren en pequeñas cantidades, hace que las reacciones bioquímicas se encuentren más expuestas a los efectos de las fuentes de variabilidad. Por lo tanto, como resultado se genera una heterogeneidad en la expresión génica de las células de una población celular isogénica (Barizien, 2019). De ahí que la variabilidad celular se encuentre de forma omnipresente en la mayoría de los sistemas biológicos, en los cuales influye en el modo en que una célula de una población celular responda de forma distinta ante un mismo estímulo. Esta variación en la respuesta de un sistema biológico se observa comúnmente en contextos como la resistencia a drogas contra el cáncer, la expresión génica, o la infección microbiana (Loos, 2019) (Filippi, 2016).

4.2.3 Modelado de redes de reacciones bioquímicas

El modelado de una reacción química conlleva entender los factores involucrados y expresarlos en la notación correcta. Comúnmente los procesos biológicos se generan a partir de la interrelación de más de una reacción bioquímica. Por ende, la mejor forma de expresar dichos procesos es mediante la definición de una serie de ecuaciones que caracterizan el comportamiento de la red de reacciones bioquímicas. La complejidad de una determinada red depende del nivel de detalle con el cual se va a modelar un sistema biológico, y desde luego, del conocimiento experimental disponible sobre el sistema bajo estudio. Otro aspecto importante al momento de modelar el comportamiento de una población celular es el proceso de adquisición de los datos experimentales. Dicho proceso usualmente parte de establecer un estímulo de entrada con el cual se incita y se cambia el entorno extracelular de una población celular; para determinar posteriormente el perfil de la respuesta en el tiempo tanto de la población como de cada uno de los individuos

que la componen (González, 2014). Por ejemplo, una estrategia usada con frecuencia consiste en usar promotores inducibles para modular la expresión de un gen de interés. Luego, la expresión génica puede ser medida a través de microarreglos o PCR cuantitativa en tiempo real, y observada a nivel celular mediante la combinación de proteínas indicadoras de fluorescencia y técnicas de microscopía de fluorescencia o citometría de flujo (Uhlendorf, 2011) (Zechner, 2012).

En la Figura 2 se muestran los pasos típicos de abstracción en el modelado matemático. Como se puede observar, se parte de un sistema biológico, en el cual se producen miles de diferentes proteínas, de las cuales comúnmente son de interés aquellas proteínas que por ejemplo se encuentran etiquetadas con un marcador de fluorescencia. De este modo, se mide el nivel de brillo de las células en función de la concentración de dicho marcador. Luego, en un modelo mental simplificado, se asume que las proteínas de interés interaccionan con ciertas moléculas que se difunden libremente en el interior de la célula; mientras que otras sustancias no son tenidas en cuenta con el fin de conservar la simplicidad del modelo. Como el modelo es una representación abstracta de los procesos que ocurren en un sistema biológico específico, una determinada red de reacciones bioquímicas puede ser representada mediante un bosquejo gráfico, donde las proteínas o metabolitos son representados por nodos, mientras que las reacciones son representadas por flechas. Partiendo de esta representación es posible describir la red de reacciones bioquímicas mediante un sistema de ecuaciones diferenciales, a partir del cual se puede cuantificar la razón de cambio en la producción de una molécula y, por ende, simular y predecir la dinámica en el tiempo de la concentración de una determinada proteína. Si los resultados no coinciden con los datos experimentales, significa que el modelo es defectuoso o se encuentra muy simplificado (Klipp, 2016). Este último corresponde a uno de los grandes retos para la biología de sistemas, el cual consiste en encontrar el grado correcto de simplificación de un sistema biológico.

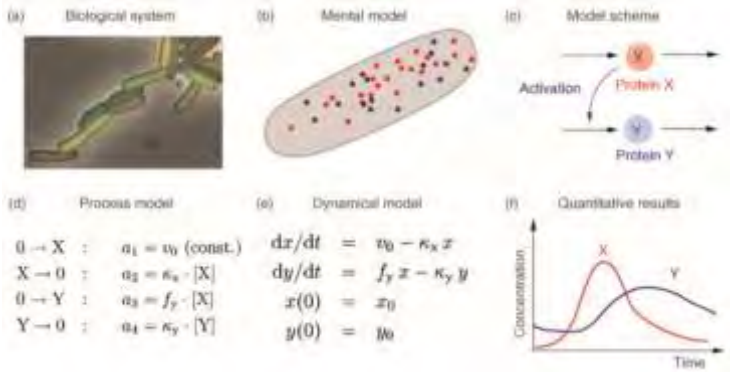

Figura 2. **Pasos típicos de abstracción en el modelado matemático.**

De *Systems Biology: A textbook* (2ª ed., p. 5), por E. Klipp, Alemania, Wiley-VCH Verlag GmBH & Co. KGaA. 2016.

Para modelar un proceso biológico frecuentemente se parte de los datos experimentales registrados de un sistema biológico. Dichos datos consisten en mediciones repetidas tomadas sobre distintos individuos de una determinada población, y constituyen lo que se denomina como “enfoque poblacional”, el cual es de gran relevancia para la caracterización y modelado de un sistema biológico. Posteriormente, se debe determinar la expresión que refleja el comportamiento típico de la población. Por ejemplo, para modelar el aumento de peso de una población de ratones, se puede emplear un modelo de crecimiento que describa el incremento de peso de los ratones en función del tiempo, y un modelo estadístico que describa la variabilidad del peso entre los individuos de la población. Para el anterior ejemplo, se puede considerar un modelo de crecimiento simple de la forma:

$$f(t) = a + b(1 - e^{-kt}) \qquad (1)$$

La Ecuación 1 se puede adaptar a las curvas de crecimiento observadas experimentalmente a partir de un conjunto de parámetros (a, b, k), los cuales pueden ser interpretados biológicamente de la siguiente manera: a como el peso de un ratón en el instante de tiempo $t = 0$, b como el peso máximo cuando el tiempo tiende a infinito, y k como una razón de cambio del peso de la población o de un individuo de esta. Una vez establecido el modelo, las variables y los parámetros que lo caracterizan, desde el enfoque poblacional se determina la curva del comportamiento típico de la población, la cual es definida por un conjunto de parámetros poblacionales $(a_{pop}, b_{pop}, k_{pop})$. Por su parte, el comportamiento de la curva de cada individuo de la población depende a su vez de su propio conjunto de parámetros individuales (a_i, b_i, k_i), los cuales pueden ser vistos como una distribución de probabilidad cuyo valor varía aleatoriamente de un individuo a otro. (Lavielle, 2014). La Figura 3 muestra el perfil de la expresión de una proteína fluorescente en una población de células de levadura. La línea azul oscuro representa el comportamiento típico de la población, mientras que las líneas azul claro representan las diferentes respuestas de cada una de las células y su distribución aleatoria alrededor de la respuesta típica de la población.

En la construcción de un modelo se deben definir las características esenciales de un determinado sistema biológico de interés, es decir, determinar los aspectos importantes que necesitan ser incluidos en el modelo, como las cantidades que pueden ser medidas o controladas, y las condiciones del entorno y limitaciones que impone el sistema. Estos aspectos se integran a un modelo comúnmente en forma de variables, parámetros o constantes. Con base en las características del sistema se pueden construir diferentes tipos de modelos, los cuales difieren en el nivel de resolución que logran y las suposiciones que conllevan (Stan, 2017). Una clasificación amplia de estos métodos separa los modelos resultantes en modelos basados en ecuaciones diferenciales ordinarias, modelos estocásticos y modelos de efectos mixtos. A continuación, se describen algunos tipos de modelos usados

comúnmente para modelar la expresión génica, sus correspondientes definiciones matemáticas y técnicas para la estimación de sus parámetros.

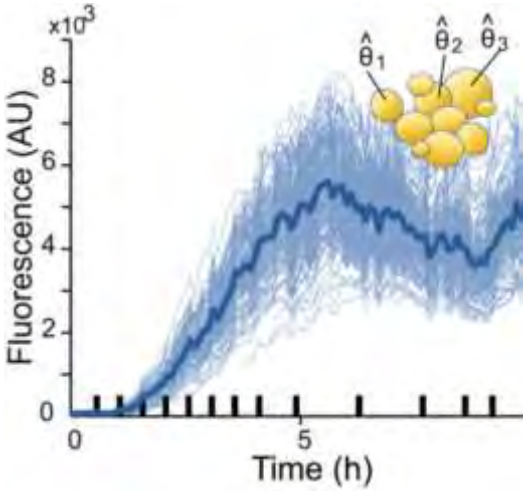

***Figura 3*. Respuesta individual y poblacional en los niveles de expresión de una proteína.**

De "What Population Reveal about Individual Cell Identity: Single Cell Parameter Estimation of Models of Gene Expression in Yeast" (p. 3), por A. Llamosi et al., 2016.

4.2.3.1 Modelos determinísticos

Los procesos biológicos, tales como la transcripción y la traducción, son procesos compuestos de distintas reacciones bioquímicas que pueden ser descritas de forma práctica mediante el uso de un conjunto de ecuaciones diferenciales ordinarias (ODEs). Dichas ecuaciones cuantifican la sucesión de estados, usualmente la concentración de una determinada especie, adoptados por una red de reacciones bioquímicas a través del tiempo (El Samad, 2005). Los modelos que utilizan ODEs para simular poblaciones celulares pueden denominarse modelos de Célula Promedio. Este tipo de modelo parte del hecho de que los procesos bioquímicos que involucran grandes cantidades de diferentes especies moleculares tienden a presentar un comportamiento promedio que hasta cierto punto puede considerarse determinista. Así, es posible obtener la caracterización de la respuesta dinámica de una red de reacciones bioquímicas en forma de la respuesta celular media mediante los parámetros cinéticos θ, la entrada u, y los estados iniciales x_0 (Almquist, 2017). Adicionalmente, en el modelo se introduce variabilidad en la respuesta del sistema en forma de error del dispositivo de medición. El valor del error de medición e puede ser asumido como una distribución Gaussiana $\eta(t)$, y se encuentra determinado por un término aditivo e_a; el cual es una constante independiente de la variable medida,

y un término multiplicativo e_b; el cual es proporcional a la variable medida. Por ende, tomando un valor diferente de $\eta(t)$ para cada célula, es posible considerar la variabilidad celular en el sistema; sin embargo, cabe resaltar que la población sigue conservando el mismo conjunto de parámetros $\{\theta, e\}$ para toda la población. Así, en un grupo de N células, la respuesta individual puede ser descrita como (González, 2014):

$$y_i(t) = f(t, u, x_0, \theta) + h(f(t, u, x_0, \theta), e)\eta_i(t) \quad (2)$$

4.2.3.2 Modelos estocásticos

En la naturaleza el número de una especie molecular se encuentra de forma discreta, las reacciones ocurren en grupos de pocas moléculas, y sus cantidades cambian igualmente de forma entera discreta (El Samad, 2005). Lo anterior causa que la dinámica de un sistema sea susceptible al ruido, y por lo tanto otorga un carácter estocástico a la respuesta del sistema. Los modelos estocásticos toman en cuenta lo anterior para describir la evolución en el tiempo de las reacciones bioquímicas a nivel molecular. Comúnmente, desde un punto de vista de los modelos estocásticos, la dinámica de un sistema evoluciona conforme a lo que se denomina *ecuación química maestra* (CME) (Cao, 2004):

$$\frac{dp(x,t)}{dt} = \sum_{r=1}^{R} a_r(x - v_r)p(x - v_r, t) - a_r(x)p(x, t) \quad (3)$$

La Ecuación 3 es un enfoque estándar que considera una red de reacciones bioquímicas que implica S especies distintas sometidas a R reacciones (González, 2013). Donde cuyos posibles estados $x = (x_1, \dots, x_S) \in N^S$ corresponden al número de moléculas de cada una de las especies $x_S(t) \in N$ en un instante de tiempo t (Ruess, 2013). La dinámica de las reacciones se encuentra caracterizada por la función de propensión $a_r(x)$ y por el vector de cambio de estado $v_{r,s} \in Z$, donde $v_r = (v_{r,1}, \dots, v_{r,S})$. El resultado de la CME es la distribución de probabilidad $p(x, t)$ correspondiente a cada especie del sistema en un determinado momento (González, 2014).

4.2.3.3 Modelos de efectos mixtos

Los modelos de efectos mixtos (ME) pueden ser usados como una alternativa para añadir una fuente de variabilidad adicional a la respuesta de los individuos de un sistema. Esto puede ser logrado considerando el caso en el cual los parámetros cinéticos varían de un individuo a otro. En los modelos ME, los parámetros se subdividen en *efectos fijos* μ, el cual es un vector que representa los parámetros que permanecen constantes en la población; y en *efectos aleatorios* $b_i \sim N(0, \Omega)$, que determina cómo los parámetros de un individuo difieren estocásticamente de los parámetros típicos de la población. La función $d(\cdot)$ establece una relación entre los

efectos fijos y los efectos aleatorios con el vector de covariables a_i; que puede ser escrita como un conjunto de parámetros compuestos para un individuo i como $\psi_i = d(a_i, \mu, b_i)$ (Wang, 2019). Otra forma de añadir no linealidad al modelo puede provenir de asumir que la variabilidad en el error varía de un individuo a otro y posiblemente con el tiempo. De este modo, los modelos ME pueden partir de una representación matemática de un modelo como el de la Ecuación 2, para expresar la variabilidad en la respuesta de los individuos de una población como (Lavielle, 2014):

$$y_i(t) = f(t, u, x_0, \psi_i) + h(f(t, u, x_0, \psi_i), e_i)\eta_i(t) \quad (4)$$

4.2.3.4 Estimación de parámetros de un modelo

Uno de los métodos utilizados para la estimación de parámetros de un modelo es conocido como la divergencia de Kullback-Leibler (KLD). Este método es una medida estadística que cuantifica que tan semejante es una distribución de probabilidad $p = \{p_j\}$ a una distribución modelo $q = \{q_j\}$ (Singh, 2013). Definida en T, con $j = 1, \dots, T$, la minimización de la KLD para obtener la estimación del mejor conjunto de parámetros θ se encuentra dada por (González, 2014):

$$\hat{\theta} = \arg\min\left(\frac{1}{T}\sum_{j=1}^{T} KLD(p||q)\right) \quad (5)$$

Por otra parte, se encuentran los métodos basados en la estimación máxima de la verosimilitud (MLE), el cual comúnmente es usado para ajustar un modelo a un conjunto de datos observados. La verosimilitud del parámetro θ dado el dato observado Y, se representa por la función L; por lo tanto, como tal es una función de Y que permite obtener la probabilidad de un determinado parámetro para un conjunto fijo de datos observados (Myung, 2003). La MLE permite determinar la estimación de los parámetros como (González, 2014):

$$\hat{\theta} = \arg\max L(\theta, Y) \quad (6)$$

5. ANÁLISIS Y DEFINICIÓN DE REQUERIMIENTOS

En este capítulo se da solución al primer objetivo específico del proyecto, el cual consiste en realizar un análisis de requerimientos para el software planteado, basándose en las necesidades de los posibles usuarios del sistema. El análisis y la definición de los requerimientos se llevó a cabo en base a la información extraída de la literatura sobre el modelado de procesos biológicos en poblaciones celulares, algoritmos implementados y la exploración del contenido del curso de Biología Computacional de la Universidad Autónoma de Occidente que se llevó a cabo en el periodo 2018-3.

5.1 IDENTIFICACIÓN DE LAS NECESIDADES

La identificación de las necesidades partió de reconocer la utilidad que tiene la biología de sistemas para la investigación biomédica, ya que esta ayuda a estudiantes e investigadores a familiarizarse con las representaciones matemáticas de los sistemas biológicos. Adicionalmente, junto con el profesor de la asignatura de Biología Computacional, se buscó relacionar los conceptos teóricos con las habilidades prácticas necesarias para modelar procesos biológicos a través de una herramienta informática. Por lo tanto, inicialmente se planteó que la herramienta debería contribuir a desarrollar las siguientes competencias en los usuarios:

- Análisis detallado de los aspectos dinámicos y matemáticos de un proceso físico, químico o biológico.

- Expresión matemática de los modelos de sistemas biológicos individuales y poblacionales.

- Simulación de los diferentes comportamientos e identificación de los parámetros que mejor describen un sistema.

Durante el proceso de identificación de necesidades se analizaron aproximadamente 50 documentos entre artículos de investigación, artículos de revisión y libros. De dichos documentos se concluyó que los más relevantes eran los siguientes:

- Systems Biology: a textbook. Klipp, E., Liebermeister, W., Wierling, C., y Kowald, A.

- Modeling and simulating chemical reactions. Higham, D. J.

- Modeling of biological processes in cell populations. Gonzalez, A. M.

- Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference. Gonzalez, A. M., Uhlendorf, J., Schaul, J., Cinquemani, E., Batt, G., y Ferrari-Trecate, G.

- Numerical solution of stochastic models of biochemical kinetics. Ilie, S., Enright, W. H., y Jackson, K. R.

- Mixed effects models for the population approach: models, tasks, methods, and tools. Lavielle, M.

- Construction and control analysis of biochemical network models. Liebermeister, W.

- Mathematical modeling of variability in intracellular signaling. Loos, C., y Hasenauer, J.

- Mathematical modelling, analysis, and numerical simulation for a general class of gene regulatory networks. Pájaro Diéguez, M.

- Quantifying intrinsic and extrinsic variability in stochastic gene expression models. Singh, A., y Soltani, M.

A partir de estas fuentes se estableció que los temas más apropiados para ayudar en el proceso de aprendizaje mediante una herramienta inicialmente conllevan el proceso de definición de un sistema biológico, incluyendo todos los datos biológicos necesarios para obtener el modelo del sistema. Posteriormente, las fuentes de variabilidad que afectan un determinado proceso se tuvieron en cuenta en diferentes algoritmos de simulación enfocados de forma determinística, estocástica, y efectos mixtos, y que emulan las variaciones en los sistemas biológicos debidas a ruido de medición, estocasticidad intrínseca y variabilidad extrínseca. Por último, se revisaron los modelos mediante los cuales se obtienen los parámetros característicos del sistema, para lo cual se siguió un enfoque centrado en las fuentes

de variabilidad de tal manera que se establezca un modelo de inferencia para cada fuente simulada.

Se investigó sobre los distintos softwares o herramientas disponibles actualmente tales como los mencionados en los antecedentes. Se concluyó que estos eran similares a lo que se pretendía implementar en la herramienta. Por ejemplo, *Lixoft* ha desarrollado una serie de herramientas para el modelado y simulación de modelos dirigidos al desarrollo de medicamentos. Sus productos ofrecen soluciones potentes y fáciles para el análisis poblacional en ensayos clínicos y la personalización de tratamientos. Aunque, presenta herramientas de gran calidad y robustez como *MONOLIX*, esta se encuentra enfocada netamente a modelos no lineales de efectos mixtos y su uso puede llegar a ser difícil debido a la complejidad de sus características y los análisis que ejecuta requiere de un mayor conocimiento sobre el tema. Adicionalmente, el uso de dichas herramientas se encuentra restringido a licencias que requieren de autorización previa (Lixoft, 2021). Por su parte, para la solución numérica de modelos estocásticos de sistemas biológicos se encontraron varias librerías en *Python* como *gillespy2*, la cual ofrece un enfoque orientado a objetos para la creación de modelos matemáticos y una variedad de métodos para ejecutar simulaciones en el tiempo de dichos modelos. Estas librerías brindan una solución rápida a un modelo estocástico; sin embargo, aun requieren de cierto grado de conocimiento en programación para poder ser utilizadas e integradas a un problema determinado (Gillespy2, 2021). En (Schnoerr, 2015) describen un nuevo paquete software llamada *MOCA*, el cual permite el análisis numérico automático de varios métodos de aproximación de cierre de momentos en una interfaz gráfica. Los métodos de aproximación que contiene esta herramienta son de gran valor a nivel de la inferencia con modelos estocásticos. *MOCA* se encuentra implementada en *Matlab*, lo cual restringe su uso a poseer una licencia de dicho software o la versión exacta de la implementación. Para el desarrollo de la herramienta se trató de integrar las virtudes presentes en las herramientas mencionadas y proponer otras características que le otorgaron valor agregado al producto final obtenido en la ejecución de este proyecto.

De acuerdo con la problemática planteada y en función de la revisión bibliográfica realizada, una herramienta informática que permita comprender los conceptos básicos requeridos para modelar la expresión génica de poblaciones celulares debe:

- Expresar matemáticamente las ecuaciones diferenciales y la matriz estequiométrica que determinan la respuesta de un proceso biológico.

- Simular la respuesta de un determinado proceso biológico a partir de diferentes fuentes de variabilidad.

- Personalizar las características de una simulación a partir de las características experimentales de un sistema biológico.

- Permitir la visualización de la respuesta de expresión génica a nivel unicelular y multicelular.

- Caracterizar la respuesta de un sistema biológico a través de diferentes modelos que permitan la inferencia de los parámetros del sistema.

- Procesar la simulación de un sistema en una cantidad de tiempo adecuada.

- Interactuar con el usuario a través de una interfaz gráfica amigable, con uno o más elementos didácticos que contribuyan a la enseñanza del modelado de sistemas biológicos.

- Brindar acceso constante a los usuarios.

5.2 ESTABLECIMIENTO DE ESPECIFICACIONES

A partir de las necesidades identificadas se plantearon las funciones y especificaciones que guiaron el desarrollo de la herramienta y que tienen mayor relevancia en el proceso de aprendizaje de las técnicas de modelado y simulación de procesos biológicos.

Aunque se estableció una metodología de tipo cascada, la herramienta se vio sujeta a diferentes modificaciones. Por lo tanto, a parte del proceso de identificación inicial, hubo un desarrollo iterativo de los requerimientos, en el que se generaron varias versiones de la herramienta. De este modo se definieron los requerimientos que se muestran a continuación, de acuerdo con lo que se consideró que el usuario podría necesitar.

La tabla 1 muestra las funciones que se establecieron para el desarrollo de la herramienta. A partir de estas se plantearon las especificaciones que se observan en la tabla 2. Dichas especificaciones reflejan cada uno de los elementos que se implementaron en la herramienta y que son pertinentes para la enseñanza académica de conceptos básicos de la biología computacional. Igualmente, a partir de las especificaciones establecidas se empieza a abordar la forma en la que se interactuará con la herramienta.

Tabla 1.

Funciones de la herramienta.

FUNCIONES
Simulación y modelado de los procesos celulares desde diferentes enfoques; entre los que se abarcan los deterministas y estocásticos.
Visualización detallada de la evolución en el tiempo de la respuesta individual y poblacional de las especies moleculares.
Simulación de los procesos biológicos en poblaciones celulares y periodos largos de tiempo.
Definición de estímulos de entrada que intervengan en la respuesta de un determinado sistema biológico.
Cálculo matemático de las ecuaciones diferenciales que modela las diferentes especies moleculares que intervienen en un proceso biológico.
Programación de la simulación de un sistema biológico desde la interfaz gráfica.

Tabla 2.

Especificaciones de la herramienta.

ESPECIFICACIONES
Solicita datos de entrada como especies moleculares, parámetros cinéticos, y matriz estequiométrica de reactivos y productos.
Definición de inicio y duración de pulsos de estímulo de entrada del sistema en forma de escalones o modelo.
Variación de configuración de simulación mediante selección de duración del experimento, concentración inicial de las especies moleculares, y número de células.
Selección del tipo de variabilidad tal que; sea de origen de ruido de medición, estocasticidad intrínseca o extrínseca, y los factores que lo definen.
Definición de la especie molecular a graficar.
Visualización de la evolución de las especies moleculares y el estímulo de entrada en gráficas de Concentración vs Tiempo.
Solicita parámetros iniciales mediante los cuales se realiza la inferencia de los parámetros que caracterizan un sistema.
Selección del tipo de modelo bajo el cual se realizará la inferencia de parámetros, y los factores que lo definen.
Distinción entre las diferentes etapas de simulación y modelado.
Implementación de la herramienta en el lenguaje de programación de código abierto Python.
Disponibilidad de la herramienta en un repositorio Online.

Las especificaciones que se muestran en la Tabla 2, son producto de las características que se fueron analizando a lo largo del desarrollo del proyecto, con el fin de que los requerimientos fueran acordes a las necesidades identificadas y a las restricciones encontradas durante el proceso. Adicionalmente, se consideró la posibilidad de implementar un método por el cual se pudieran guardar y cargar datos simulados de un sistema, y una característica relacionada a la inferencia de parámetros en cuanto a la evolución de los parámetros a lo largo de las iteraciones ejecutadas en dicho proceso.

6. DISEÑO DE LA HERRAMIENTA

Este capítulo plantea el proceso de diseño de la herramienta; por lo tanto, tal como se estableció en el segundo objetivo del proyecto, se desarrolló el modelo conceptual de la herramienta. Este modelo conceptual tiene en cuenta los requerimientos encontrados en el capítulo anterior, así como las ventajas y restricciones dadas por los modelos matemáticos actualmente utilizados para simular los procesos biológicos relevantes para el proyecto.

6.1 DESCRIPCIÓN DE LA ESTRUCTURA GLOBAL

La arquitectura de la herramienta tiene en cuenta los requerimientos descritos anteriormente, y sigue una estructura basada en el paradigma comúnmente usado en la Biología de Sistemas para el descubrimiento de conocimiento, tal como se observa en la Figura 4. Este paradigma consiste en un proceso cíclico dividido en varias etapas que pueden ser de carácter experimental o computacional. Partiendo desde un punto de vista experimental, este proceso inicia en una etapa de diseño experimental que establece las condiciones bajo las cuales se realizara un experimento *in vitro* o *in vivo* de un determinado proceso biológico. Definidas las condiciones experimentales, se valida el modelo propuesto y se procede a la obtención de la respuesta de expresión génica de una población celular. La respuesta de una población celular puede ser registrada a través de procedimientos de adquisición de datos, tales como los que conllevan fluoroscopia o citometría de flujo; técnicas mediante las que es posible obtener imágenes de dicha respuesta. El procesamiento de las imágenes obtenidas usualmente conlleva la segmentación y seguimiento en el tiempo del perfil de respuesta de cada una de las células. Las etapas experimentales concluyen en el análisis de las observaciones obtenidas, desde donde es posible concebir nuevos conocimientos a nivel biológico, clínico, farmacológico, etc. Parte del conocimiento obtenido se ve reflejado en forma de entendimiento referente a como interactúan las reacciones y especies moleculares que definen el comportamiento de un proceso biológico específico.

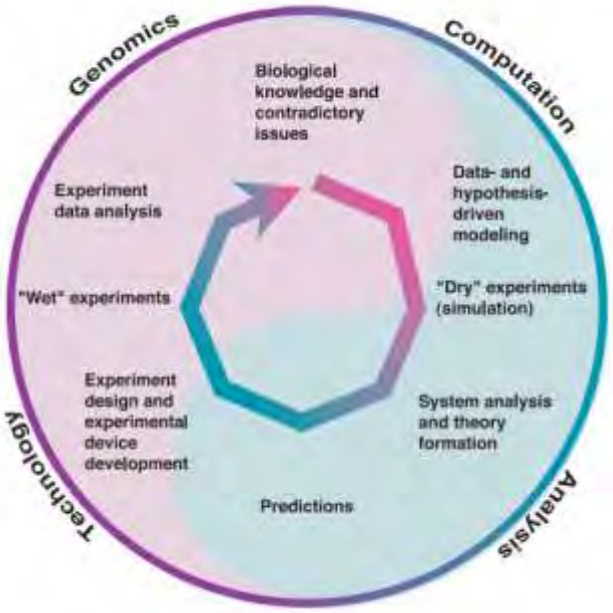

Figura 4. **Paradigma típico en Biología de Sistemas.**

De "Systems Biology: A Brief Overview" (p. 2), por H. Kitano, 2002.

Por su parte, las etapas de carácter computacional toman el conocimiento adquirido experimentalmente para proponer un modelo de un determinado proceso biológico. Desde este punto de vista, se trata de obtener la respuesta simulada que más se aproxime a las observaciones experimentales. Por lo tanto, se establecen diferentes tipos de modelos bajo los cuales se realizará la caracterización del sistema y su posterior simulación. La selección del tipo de modelo se realiza de forma intuitiva dependiendo de que fuente de variabilidad se sospecha que se encuentra regido el proceso biológico bajo cuestión. De este modo, se obtienen de nuevo perfiles de la respuesta de la población del sistema biológico. Estos datos pueden ser manipulados computacionalmente, visualizados y analizados para ser comparados con las observaciones experimentales. Partiendo de las comparaciones entre las observaciones experimentales y las simuladas es posible establecer conclusiones que permitan comprobar o refutar hipótesis; de las cuales se puede obtener mayor conocimiento de un proceso biológico y/o mejorar las condiciones de diseño de un procedimiento experimental (Kitano, 2002).

Para la estructura global de la herramienta se tomó como eje central principalmente el eje computacional del paradigma descrito previamente. Sin embargo, también se trató de emular ciertos aspectos que hacen parte del eje experimental; específicamente, se realizan simulaciones a partir de algoritmos que aproximan su respuesta a lo que correspondería a las observaciones experimentales. Lo anterior

se debe al enfoque educacional de la herramienta dirigido a estudiantes o investigadores que deseen introducirse en el campo de la Biología de Sistemas.

En la Figura 5 se presenta de forma general la estructura planteada para la herramienta. Como se puede observar, la interacción con el usuario inicia definiendo las entradas del sistema. Las características de dichos datos corresponden a la información necesaria para la simulación de los modelos implementados. La simulación del sistema biológico toma lugar en la primera etapa de funcionamiento de la herramienta, y a partir de esta se obtiene la respuesta del sistema. La etapa posterior toma como entrada las observaciones simuladas y datos iniciales, entre los cuales se encuentran los parámetros cinéticos o suposiciones iniciales, desde los que se inicia el modelado del sistema para realizar la inferencia de parámetros "reales" que caracterizan su respuesta.

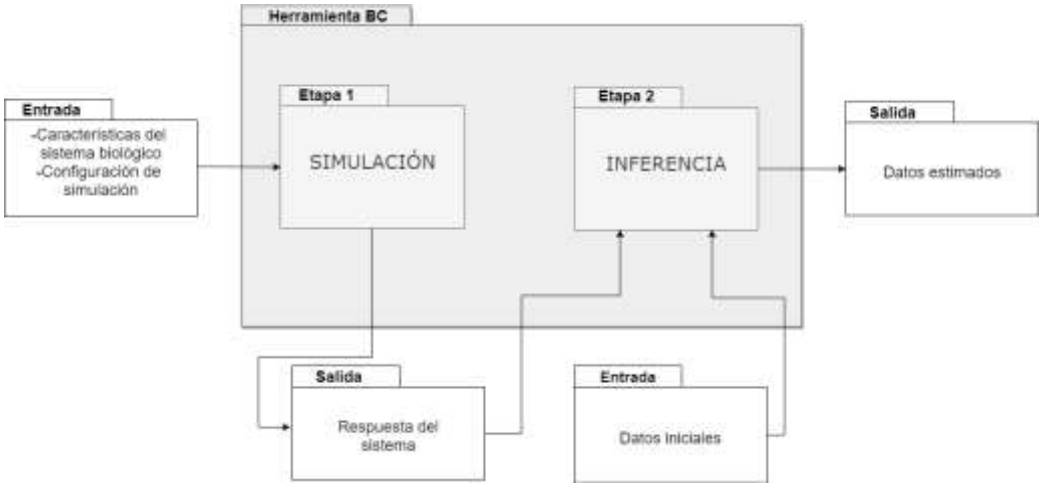

Figura 5. **Estructura global de la herramienta.**

De forma general, las etapas de simulación e inferencia se encuentran compuestas tal como se observa en las Figuras 6 (A) y 6 (B) respectivamente. La etapa de simulación inicia definiendo las características del sistema y la configuración de la simulación a realizar. A partir de esta información, y del estímulo de entrada, la herramienta define el sistema de ecuaciones diferenciales que representan el proceso biológico de interés. Una vez definido el sistema, es posible simular su respuesta desde diferentes fuentes de variabilidad. Como resultado se obtienen los perfiles de las respuestas individuales y se calculan datos estadísticos que permiten obtener la respuesta poblacional del sistema.

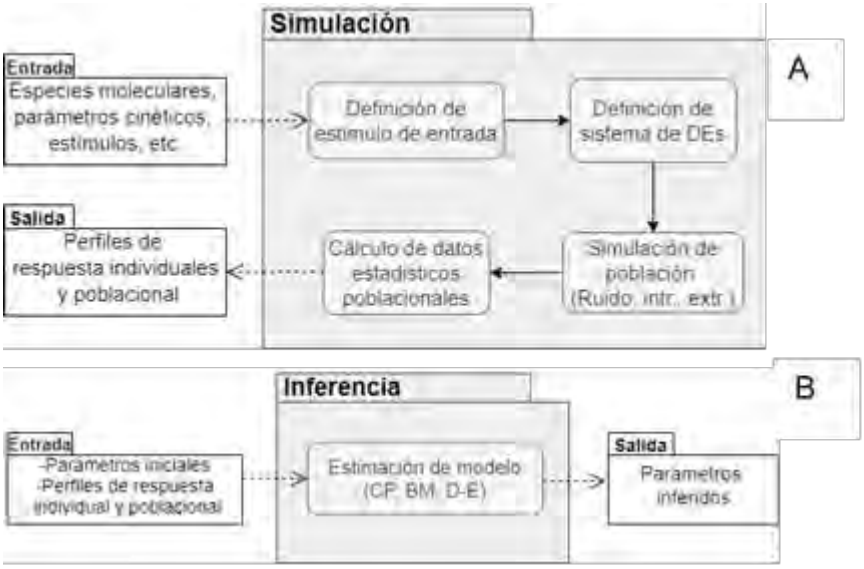

***Figura 6*. A) Estructura general de la etapa de simulación. B) Estructura general de la etapa de modelado.**

Por su parte, la etapa de inferencia inicia definiendo parámetros iniciales y tomando las observaciones simuladas, para seleccionar el tipo de modelo mediante el cual se va a realizar la caracterización del sistema e inferir los parámetros que definen la respuesta de la población.

6.2 DEFINICIÓN DE LAS UNIDADES QUE COMPONEN LA HERRAMIENTA

6.2.1 Etapa de simulación

La Figura 7 muestra de forma detallada el flujo de funcionamiento y los procesos que se llevan a cabo para completar la simulación de un sistema biológico. A continuación, se definen cada una de las etapas que intervienen en dicho proceso:

6.2.1.1 Características del sistema biológico

En este punto se establecen los aspectos base que van a definir la estructura del proceso a simular. Para definir dicha estructura se debe establecer:

- *Especies moleculares*: Diferentes tipos de moléculas involucradas en un proceso. Estas moléculas pueden tomar parte en una o más reacciones químicas.

- *Parámetros cinéticos*: Corresponden a las reacciones que se producen en el proceso, y determinan la probabilidad de que cada reacción tome lugar en un intervalo de tiempo infinitesimal.

- *Matriz de reactivos*: Representa la matriz de estados de las especies moleculares que actúan como reactivos en la red de reacciones.

- *Matriz de productos*: Representa la matriz de estados de las especies moleculares que actúan como productos en la red de reacciones.

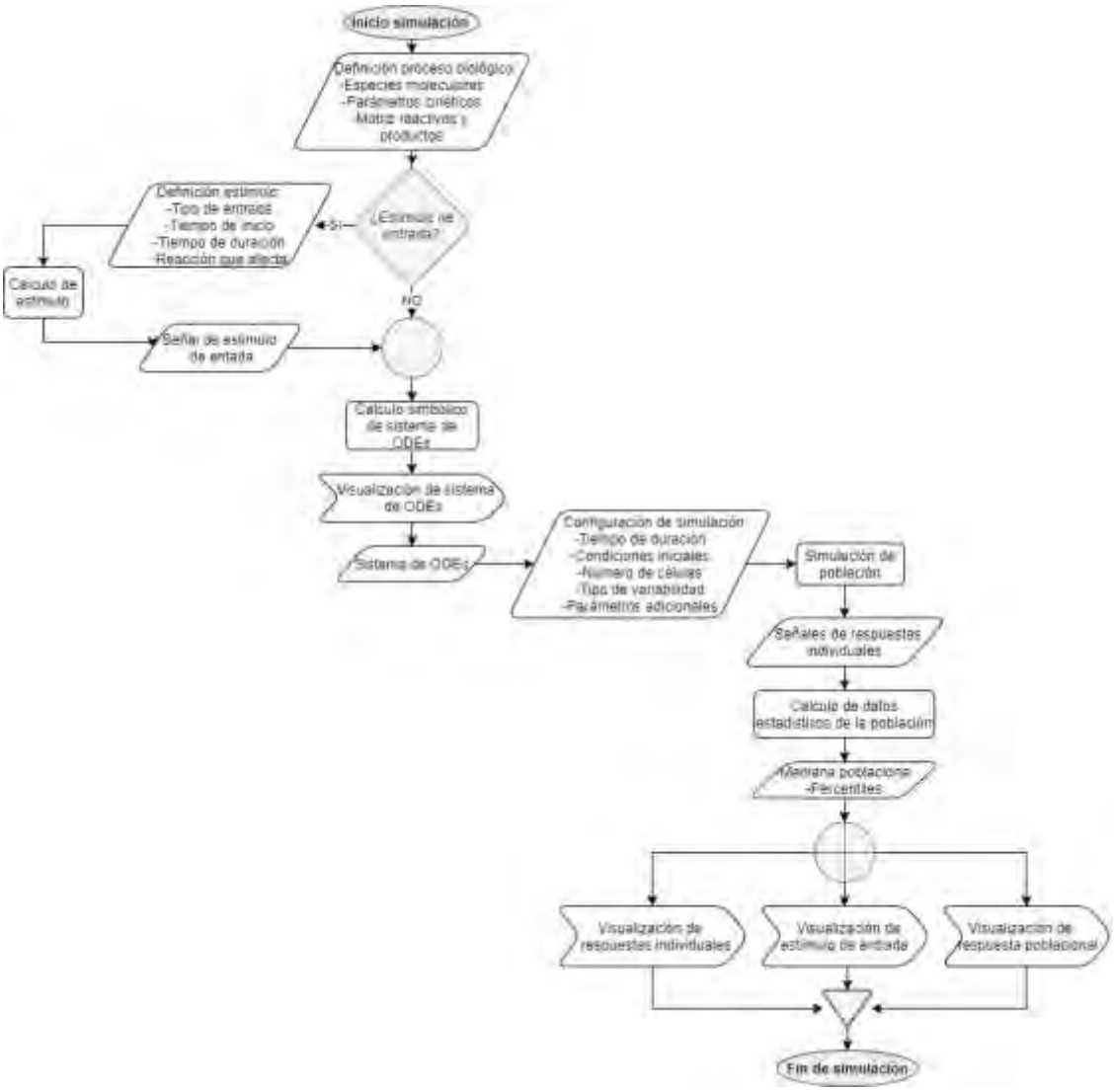

Figura 7. **Diagrama de funcionamiento de la etapa de simulación.**

6.2.1.2 Definición de estímulo de entrada

Esta unidad puede ser opcional, dependiendo de si el sistema se encuentra influenciado por un estímulo de entrada. En el caso de presentarse uno o varios estímulos de entrada, se debe definir tanto el tiempo de inicio como el tiempo de duración de cada estímulo. Además, se debe definir el tipo de estímulo y la reacción afectada por el mismo.

6.2.1.3 Definición de sistema de ecuaciones diferenciales

La definición del sistema de ecuaciones diferenciales toma la información establecida previamente, y calcula el modelo estructural del proceso biológico. Este sistema de ecuaciones es calculado de forma simbólica; es decir, se obtienen las expresiones matemáticas que definen la respuesta individual en el sistema. La definición de dicho sistema constituye un punto de partida para que el usuario pueda comprender como surge la respuesta de cada especie molecular a partir de su interacción en cada reacción bioquímica que toma lugar en el proceso biológico definido.

6.2.1.4 Simulación de población

Este punto comprende el proceso mediante el cual se obtienen las respuestas de los individuos que componen el sistema biológico. Por lo tanto, requiere que se establezcan los aspectos que configuran la simulación. Entre la información requerida se encuentran el tiempo de duración del experimento, el número de individuos (células), la concentración inicial de las especies involucradas, y el tipo de variabilidad junto con sus respectivos factores. Este último corresponde a uno de los algoritmos implementados para simular ruido de medición (aditivo y multiplicativo), estocasticidad intrínseca, y variabilidad extrínseca (matriz de covarianza).

6.2.1.5 Cálculo de datos estadísticos poblacionales

Los datos estadístico-poblacionales son calculados a partir del total o de un segmento de percentil de la distribución de las respuestas individuales. La implementación de la herramienta se limita a calcular la mediana poblacional, y la respuesta poblacional contenida entre los percentiles 97.5 y 2.5 de la distribución total.

6.2.1.6 Respuesta del sistema

Una simulación concluye con la visualización de la respuesta de las unidades anteriores. Iniciando con la respuesta obtenida inmediatamente de la unidad de simulación de población. En la gráfica obtenida se visualizará la evolución en el tiempo de la expresión de una determinada especie molecular de cada uno de los individuos simulados en el sistema biológico. Posteriormente, se visualizará la señal obtenida como estímulo de entrada, y por último se visualizará la respuesta poblacional del sistema, tal que se pueda observar cómo evoluciona el comportamiento medio y los límites entre los cuales se encuentra contenida la respuesta de la población.

6.2.2 Etapa de inferencia

La etapa de inferencia toma las observaciones simuladas para obtener los parámetros que caracterizan la respuesta del sistema biológico. En la Figura 8 se puede observar el flujo del funcionamiento de esta etapa. Las unidades que la comprenden se definen como:

6.2.2.1 Datos iniciales

La inferencia es un proceso que inicia principalmente con la definición de los valores iniciales de los parámetros que se pretenden inferir. Dichos valores se definen a partir de conocimiento previo sobre el sistema. Adicionalmente, es posible establecer el tipo de modelo que se utilizara para realizar la inferencia; así como el número de iteraciones y la especie molecular usada para realizar la estimación.

6.2.2.2 Estimación del modelo

Para la estimación del modelo, se realiza la minimización de una función de costo; cuya naturaleza depende del modelo seleccionado para realizar la inferencia. Un criterio para la selección del modelo se encuentra determinado por el tipo de variabilidad que se quiere identificar en el sistema. Los parámetros inferidos son aquellos valores óptimos a partir de los cuales el residuo de la función de costo es minimizado.

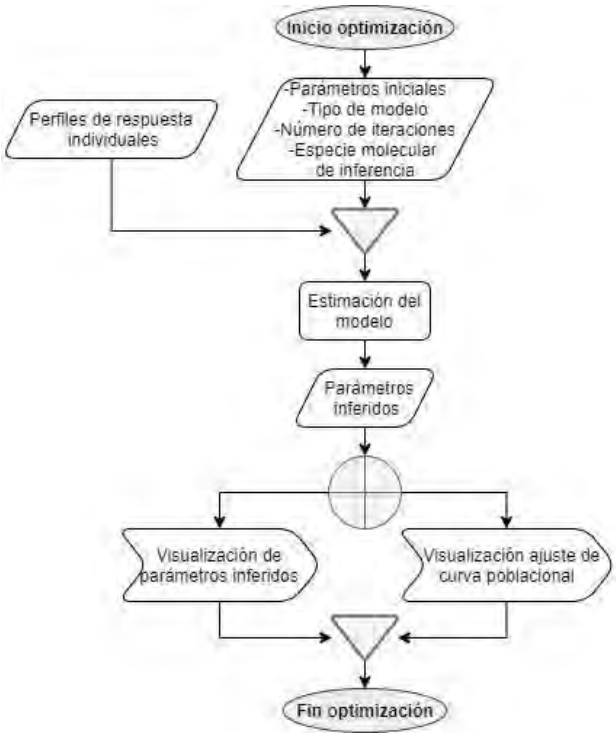

Figura 8. **Diagrama de funcionamiento de la etapa de inferencia.**

6.2.2.3 Datos inferidos

La inferencia finaliza mostrando los parámetros inferidos que caracterizan la respuesta de un determinado sistema biológico. Cabe resaltar que dichos parámetros pueden ser iguales o aproximados a los utilizados en la etapa de simulación. Sin embargo, pueden diferir si al momento de minimizar la función de costo, el algoritmo de estimación encuentra un óptimo local en el cual los parámetros inferidos, aunque distintos de los originales, permiten obtener una respuesta similar a la de las observaciones. A partir de los parámetros inferidos y de sus valores iniciales se puede graficar la respuesta poblacional del sistema y comparar ambos instantes de la etapa de inferencia.

7. IMPLEMENTACION DE LA HERRAMIENTA

A continuación, se describe el lenguaje de programación utilizado para la implementación de la herramienta teniendo en cuenta, tal como se estableció en el tercer objetivo específico del proyecto, que sea de uso libre. De igual manera, se profundiza y se explican los conceptos matemáticos tenidos en cuenta durante la implementación de cada una de las unidades y algoritmos que integran la herramienta.

7.1 IMPLEMENTACIÓN DE LAS UNIDADES

La implementación de las unidades se llevó a cabo en el lenguaje de programación *Python*, específicamente en la versión 3.7.4. Este lenguaje de programación se eligió porque es ampliamente usado en la comunidad científica debido a su versatilidad para lidiar con problemas complejos; entre los cuales, *Python* destaca en la recolección y limpieza, exploración, modelado y visualización de datos. Adicionalmente, proporciona todas las herramientas necesarias para la manipulación de grandes cantidades de datos; dado que posee bibliotecas estadísticas y numéricas robustas como *Pandas*, *Numpy*, *Matplotlib*, *Scipy*, *scikit-learn*, etc; las cuales son fáciles de usar y cuentan con un gran soporte por parte de los programadores de la comunidad pythonista. Otra razón para la elección de *Python* se debe a que es un lenguaje de programación de software libre; lo cual otorga total libertad sobre los proyectos desarrollados en este lenguaje. Lo anterior garantiza una mayor distribución y acceso a herramientas informáticas por parte de la población, especialmente en entornos educativos.

Los módulos *Python* utilizados para la implementación de la herramienta fueron los siguiente:

- ***Numpy 1.16.5***. Es una librería que permite manejar arreglos multidimensionales y matrices de gran tamaño. Brinda una gran colección de funciones matemáticas de alto nivel que permite operar sobre dichos arreglos de forma eficiente.

- ***Scipy 1.5.3.*** Es una librería enfocada a la computación científica y técnica de problemas en matemáticas, ciencia e ingeniería.

- ***Sympy 1.4***. El objetivo de esta librería es facilitar la computación simbólica de expresiones algebraicas.

- ***Matplotlib 3.3.3***. Comprende una biblioteca bastante completa para la creación de visualizaciones estáticas, animadas e interactivas.

- ***PyQt5 5.12.3***. En su versión 5, es un enlace de *Python* para el kit de herramientas de aplicaciones multiplataforma Qt. Permite la implementación fácil y rápida de interfaces gráficas.

La herramienta fue implementada usando un computador portátil con un procesador AMD A8-6410 APU @2.00 GHz con 8 GB RAM, por lo que cualquier computador con características equivalentes podría ejecutar el programa de forma adecuada. Asimismo, también se necesita tener al menos la versión 3.7.4 de Python con las correspondientes librerías instaladas.

A continuación, se describen detalladamente los principios matemáticos seguidos durante la implementación de cada una de las unidades de la herramienta. La implementación de la herramienta se encuentra disponible en el siguiente enlace de *GitHub*. Allí, se podrá acceder a los códigos correspondientes a la GUI y a cada una de las unidades individuales que componen la herramienta.

- https://github.com/Jebrayam/systemsbiology

7.1.1 Definición de sistema de ecuaciones diferenciales

Los procesos celulares se representan usualmente como sistemas de reacciones químicas. La evolución de estos sistemas puede ser modelada de forma determinística o estocástica mediante ecuaciones diferenciales (Ilie, 2009). De este modo, es posible describir la respuesta de un determinado sistema a partir del cambio de concentración de una especie molecular a través del tiempo. En primer lugar, se considera un sistema que involucra N diferentes tipos de especies moleculares $\{S_1, \dots, S_N\}$. Estas especies pueden participar en uno o más de M reacciones bioquímicas $\{R_1, \dots, R_M\}$ (Higham, 2008). Adicionalmente, se considera que el sistema se encuentra en un estado de agitación, dentro de un volumen fijo y temperatura constante. En este enfoque se ignora la información espacial y simplemente se enfoca en seguir el número de moléculas de las especies en cada instante de tiempo. La dinámica del sistema se encuentra descrita por un *vector de estado* $X(t) = [X_1(t), X_2(t), \dots, X_N(t)]$, donde $X_i(t)$ es un entero no negativo que registra el número de moléculas de la especie S_i en el instante de tiempo t. Inicialmente, la evolución del sistema parte del estado inicial de las especies; es decir, $X(t_0) = x_0$ (Lei, 2011).

Generalmente, cada reacción R_j se encuentra asociada con un vector de cambio de estado; también conocido como *vector o matriz estequiométrica* $(v_{j,1}, \dots, v_{j,N})$, cuya componente i corresponde al cambio en el número de moléculas S_i como consecuencia de la reacción j. De igual forma, cada reacción está asociada con una *función de propensión* $a_j(X(t))$, la cual expresa la probabilidad de que la reacción j tome lugar en un intervalo de tiempo $[t, t + dt)$ dado por $a_j(X(t))dt$. La matriz estequiométrica se obtiene mediante el conteo del número de moléculas de cada especie que son consumidas y producidas en una reacción. Por su parte, la función a_j tiene la forma matemática $a_j(X) = c_j h_j(X)$. Aquí, c_j es el parámetro cinético para cada reacción R_j, mientras que h_j mide el número de combinaciones distintas de moléculas reactivas R_j disponibles en el estado X (Lei, 2011).

Así, los procesos celulares evolucionan de acuerdo con un conjunto de N ecuaciones diferenciales obtenidas mediante la siguiente expresión:

$$\frac{d_{X_i}}{dt} = \sum_{j=1}^{M} v_{ji} a_j(X) \tag{7}$$

7.1.2 Entrada del sistema

En muchos procesos de expresión génica, la transición de un promotor de un estado activo a inactivo se encuentra regulada por proteínas, las cuales actúan como activadores o represores. En el caso de la activación, el activador se une al promotor inactivo para aumentar los niveles de expresión génica (Lei, 2011). El grado de activación de un determinado promotor depende principalmente de la intensidad y la frecuencia de aplicación de un estímulo específico en el ambiente celular. A nivel experimental, dicho estimulo se presenta comúnmente en forma de un *shock osmótico* liberado mediante una válvula a una cámara (ej.: cámara microfluidica). Por su parte, a nivel computacional, los pulsos de shock osmótico pueden ser representados a partir de establecer los instantes de tiempo en los cuales la válvula se abrirá, y el periodo de tiempo que permanecerá en ese estado. Cabe resaltar que se debe asegurar que los pulsos de shock osmótico no se sobrepongan entre ellos. De este modo, t_{on} representa el vector que contiene los instantes de tiempo de apertura de la válvula, y t_{dur} contiene la duración de cada pulso p. Por lo tanto, el estado de la válvula $u_v(t)$ se encuentra determinado por ambos vectores.

$$t_{on} = [t_{on1}, t_{on2}, \dots, t_{onp}] \tag{8}$$

$$t_{dur} = [t_{dur1}, t_{dur2}, \dots, t_{durp}] \tag{9}$$

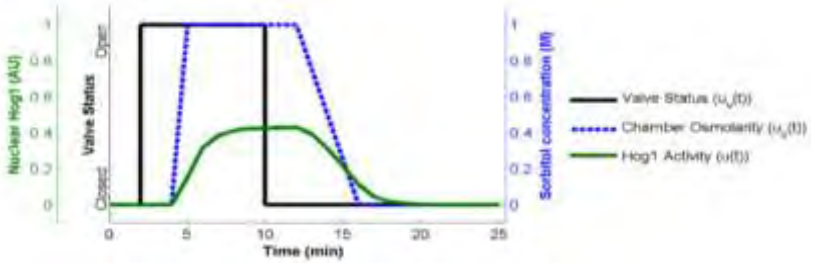

Figura 9. **Perfil del estímulo de entrada de un sistema biológico.**

De "Modeling of Biological Processes in Cell Populations" (p. 14), por A. M. Gonzalez, 2014.

Los pulsos de shock osmóticos pueden ser aplicados en forma de tren de pulsos escalón, o en su forma percibida por el sistema definida por su propio modelo estructural. Un ejemplo de este último es el caso de la ruta del glicerol hiperosmótico (HOG). La producción del glicerol se activa en respuesta a un shock osmótico, el cual produce un cambio en el ambiente celular y altera la homeostasis en el interior de las células. En consecuencia, la activación del HOG dispara procesos intracelulares en el citoplasma como el cierre de canales *Fps1p* y la activación de varias quinasas. Igualmente, este proceso desencadena la modificación en la expresión de varios genes mediada por factores de transcripción. Esta respuesta favorece la acumulación de glicerol y resulta en la restauración del balance osmótico a nivel intracelular (Uhlendorf, 2011). *In Silico*, la producción del promotor de glicerol *Hog1* puede ser creada en tres fases. Primero, la generación de los pulsos de shock osmótico. Luego, la respuesta de la cámara de microfluidica a los cambios de osmolaridad. Y, por último, la expresión de la actividad de la proteína *Hog1.* En la cámara de microfluidica, las células perciben los pulsos de shock osmótico de forma distinta a su forma en escalón unitario. El perfil de los pulsos en la cámara $u_c(t)$ presenta un retraso con respecto a la actividad $u_v(t)$. Esta relación puede ser representada por una función lineal por partes, en la cual se presenta un retardo en el pulso y un decaimiento lineal, el cual ocasiona que la concentración osmótica en la cámara disminuya exponencialmente con el tiempo. El retraso en el tiempo de la señal $u_c(t)$ da lugar a la transducción de señales, la expresión génica y la síntesis de proteínas. La Figura 9 muestra los perfiles de cada una de las señales descritas anteriormente. La actividad de la proteína *Hog1* $u_h(t)$ se puede representar de forma simple mediante un modelo que depende de un conjunto de parámetros de producción k y degradación g de la especie, y que obedece la siguiente ecuación diferencial (González, 2014):

$$\dot{u}_h(t) = k u_c(t) - g u_h(t) \qquad (10)$$

7.1.3 Variabilidad: Ruido de medición

Una de las formas más sencillas de introducir variabilidad en un modelo es teniendo en cuenta el error de medición del dispositivo de medición. Las fuentes de error varían dependiendo de las técnicas experimentales aplicadas. Por simplicidad es posible asumir que el error en las mediciones tiene una distribución Gaussiana. La intensidad del ruido está dada por un término aditivo e_a, constante e independiente de la variable de medición; y un término multiplicativo e_b, el cual es proporcional al valor de la variable medida. El modelo de ruido de medición esta dado por la Ecuación 11. Este modelo puede ser aplicado en los modelos presentados en las Ecuaciones 2, 3 y 4. Donde son multiplicados por ruido Gaussiano $\eta(\cdot)$ expresado como una distribución normal estándar $N(0,1)$. Con el fin de tener en cuenta la variabilidad, se asume que $\eta(t)$ es diferente para cada célula (González, 2014).

$$h(f(t,u,x_0,\theta),e)\eta(t) = (e_a + e_b f(t,u,x_0,\theta))\eta(t) \qquad (11)$$

7.1.4 Variabilidad: Estocasticidad intrínseca

La *ecuación química maestra* (CME) utilizada para simular reacciones bioquímicas de forma estocástica consiste en un sistema de dimensiones infinitas de ecuaciones lineales, por lo tanto, las soluciones de forma cerrada solo existen para un pequeño grupo de casos (González, 2014). En la práctica se utilizan métodos de aproximación como el *algoritmo de simulación estocástica* (SSA), el cual es un método de simulación basado en el *algoritmo de Gillespie* (AG) y es un método bien conocido por ser rigurosamente equivalente a la CME. La naturaleza estocástica de este tipo de modelo se encuentra en como el algoritmo de forma aleatoria determina el tiempo τ en el cual sucederá la siguiente reacción y el índice μ del tipo de esta.

La simulación del AG parte de establecer una matriz estequiométrica V donde se representen los reactivos y los productos de la red de reacciones de un determinado proceso bioquímico. El proceso del AG consiste en una serie de pasos que se repiten un número aleatorio de veces para cada célula en una población.

Paso 0. Inicialmente se establecen los valores de los parámetros cinéticos de las reacciones c_v $(v = 1, \dots, R)$, los valores iniciales de las especies X_i $(i = 1, \dots, S)$, el valor inicial de tiempo $t = 0$, y una variable $n = 0$ que alcanzara su valor final de forma aleatoria de acuerdo con el número de iteraciones del algoritmo.

Paso 1. La dinámica de las reacciones está dada por la función de densidad de probabilidad $p(\mu, \tau)dt$, la cual expresa la probabilidad que dado el estado de X_i en el instante de tiempo t; la siguiente reacción tomará lugar en el intervalo de tiempo infinitesimal $(t + \tau)$, y será una reacción de tipo R_μ . Así, p se aproxima mediante el cálculo de la función de propensión a_v. Dicha función expresa la probabilidad de que una reacción especifica ocurra y se determina mediante el producto entre c_v y h_v. Este último corresponde al número de combinaciones distintas disponibles en el estado $(X_1, X_2, \dots, X_S)$. Posteriormente, se calcula la probabilidad de que cualquier reacción ocurra a_o a partir de la sumatoria de los valores de a_v calculados para cada reacción.

$$a_v = h_v c_v \tag{12}$$

$$a_0 = \sum_{v=1}^{R} a_v(X) \tag{13}$$

Paso 2. En este punto se generan dos números aleatorios r_1 y r_2 que cumplan $N(0,1)$, los cuales son usados para calcular los valores de τ y μ respectivamente. Así, el instante de tiempo para que ocurra la siguiente reacción esta dado por $t + \tau$, donde τ se determina a partir de convertir el número aleatorio uniforme r_1 en un número aleatorio distribuido exponencialmente.

$$\tau = \frac{1}{a_0} \ln\left(\frac{1}{r_1}\right) \tag{14}$$

Por su parte, el valor del índice μ de la reacción que ocurrirá corresponde al entero más pequeño que satisface:

$$r_2 a_0 \leq \sum_{v=1}^{\mu} a_v \tag{15}$$

Paso 3. Obtenidos los valores de τ y μ, se incrementa el valor de t tal que $t = t + \tau$, y se ajustan los niveles de las especies tal que se refleje el efecto de la reacción R_u. Por último, se incrementa el valor de n por $n = n + 1$. De este modo, se regresa al paso 1 hasta que se cumple cierto criterio de interrupción, el cual puede estar determinado por el tiempo de simulación objetivo del proceso biológico (Gillespie, 1977).

7.1.5 Variabilidad: Extrínseca

La variabilidad extrínseca en los procesos celulares parte del hecho de tomar la influencia que tienen las diferencias morfológicas entre las células de una población sobre su respuesta de expresión génica. Para tomar en cuenta dichas diferencias,

se considera que los parámetros cinéticos varían de una célula a otra, y pueden definir estadísticamente el modelo de cambio en los parámetros. De este modo, se crea un modelo para cada individuo de la población (González, 2014). La respuesta del sistema de ecuaciones de cada individuo puede ser asumida como:

$$\dot{x}_i(t) = v(x_i, u(t), \psi_i) \tag{16}$$

Donde los parámetros se subdividen en *efectos fijos* y en *efectos aleatorios*. La combinación de ambos efectos puede ser descrita como una distribución de parámetros para un individuo i como (Wang, 2019):

$$\psi_i = d(a_i, \mu, b_i); \tag{17}$$

Donde $b_i \sim N(0, \Omega)$, μ es el vector de efectos fijos que representa los parámetros que permanecen constantes en la población. El vector b_i denota los efectos aleatorios, el cual determina como los parámetros de un individuo difieren estocásticamente de los parámetros típicos de la población. Los efectos aleatorios pueden ser establecidos a partir de una distribución Gaussiana con media 0 y covarianza Ω.

La relación entre a_i y b_i es desconocida; por ende, a_i puede ser ignorada de tal modo que se busque una correlación entre los parámetros inferidos y las observaciones independientes. Así, es posible establecer un conjunto de parámetros individuales con una distribución log-normal tal que $\varphi_i = \mu + b_i, \varphi \sim N(\mu, \Omega)$, y por lo tanto la función $d = e^{\varphi_i} = e^{\mu + b_i}$(González, 2014). Lo descrito anteriormente es aplicado en el modelo presentado en la Ecuación 4 para modelar una población celular mediante modelos de efectos mixtos.

7.1.6 Inferencia de parámetros: Célula Promedio

La inferencia de parámetros mediante el modelo de Célula Promedio se realizó a partir de la estimación de la verosimilitud máxima (MLE). Este método es comúnmente usado para ajustar un modelo a un conjunto de datos observados. Tal como se presentó en la Ecuación 6, la función L representa la verosimilitud del parámetro θ dado el dato observado Y; por lo tanto, como tal es una función de Y que permite obtener la probabilidad de un determinado parámetro para un conjunto fijo de datos observados (Myung, 2003). Para la estimación de parámetros de una población celular, Y corresponde al conjunto de observaciones registradas. Dichas observaciones se denotan como Y_{ij}, donde $i = 1, \dots, N$ y $j = 1, \dots, T$ corresponden al número total de células e instante de tiempo respectivamente. Debido a limitaciones computacionales, usualmente se realiza la MLE como la minimización de la transformación logarítmica de la función de verosimilitud. De este modo, los parámetros se estiman como (González, 2014):

$$\widehat{\theta} = \arg\min(-\log(L(\theta, Y))) \quad (18)$$

$$-\log\big(L(\theta, Y)\big) = \sum_{i=1}^{N}\sum_{j=1}^{T}\frac{1}{2}\left(\frac{Y_{ij} - f(t, u, x_0, \theta)}{h(f(t, u, x_0, \theta), e)}\right)^2 + \sum_{i=1}^{N}\sum_{j=1}^{T}\log(h(f(t, u, x_0, \theta), e)) \quad (19)$$

7.1.7 Inferencia de parámetros: KLD basada en momentos

Los métodos basados en momentos (MB) son otra alternativa para realizar aproximaciones de la CME, debido a que reducen el número de variables de estado mediante el cálculo de los momentos más relevantes de una distribución (Pájaro, 2017). A partir de MB, una alternativa para caracterizar la probabilidad $p(x,t)$, es considerar la colección de todos los momentos $M_q(t) = E[\chi^q]$, donde $q = (q_1, \dots, q_S) \in N^S$, $\chi^q = \chi_1^{q1}, \dots, \chi_S^{qS}$, y $|q| = q_1 + \cdots + q_s$ es el orden de M_q. Para las propensiones $a_r(x)$ en x y algunos momentos $L \in N \setminus \{0\}$, M denota un vector que contiene todos los momentos hasta el orden L. De este modo, se tiene que (González, 2014):

$$\frac{d}{dt}M(t) = AM(t) + B\bar{M}(t) \quad (20)$$

Donde A y B son matrices dependientes de las velocidades de reacción de la red, y $\bar{M}$ es un vector de dimensión infinita que contiene los momentos de orden estrictamente superior que L. Sin embargo, la Ecuación 20 se encuentra generalmente abierta; es decir, que no puede ser resulta debido a los momentos desconocidos $\bar{M}(t)$. Por lo tanto, dicha ecuación es aproximada mediante lo que se denomina como *cierre de momentos* (MC). El MC resulta en la solución de una ecuación similar, en la cual se reemplaza B con la función φ. Dicho cambio hace que la ecuación sea solucionable y debe ser elegida de tal manera que $M(t) \cong \widehat{M}(t)$ sobre el periodo de tiempo de interés (González, 2016):

$$\frac{d}{dt}\widehat{M}(t) = A\widehat{M}(t) + \varphi(\widehat{M}(t)) \quad (21)$$

A partir de MC, se obtienen las ecuaciones para los momentos hasta el orden L, las cuales constituyen un sistema dinámico para los momentos y el cruce de estos. Dichas ecuaciones describen la evolución en el tiempo de todos los momentos y brindan suficiente información sobre la dinámica promedio de las mediciones, lo cual hace que la identificación de parámetros sea más práctica para sistema grandes. De este modo, MC provee un medio para capturar la estocasticidad de las

reacciones de un sistema, mientras aprovecha la escalabilidad de los modelos de ecuaciones diferenciales ordinarias.

En *Zechner et al.* (2012) se muestra que las mediciones de la media y la varianza de las especies pueden ser suficiente información para determinar los parámetros que caracterizan un determinado sistema biológico. Adicionalmente, también muestra que es posible obtener de forma directa el sistema de ecuaciones diferenciales de momento basados en la población. Limitando la definición del sistema de ecuaciones a depender de los momentos de orden menor que $L = 2$. Los momentos de población aproximados resultantes se pueden obtener como (Zechner, 2012):

$$\frac{d_{x_i}}{dt} = \sum_{j=1}^{M} v_{ji} a_j(x) \tag{22}$$

$$\frac{dx_i x_k}{dt} = \sum_{j=1}^{M} (v_{ji} a_j(x) x_k \; + \; v_{jk} a_j(x) x_i \; + \; v_{ji} v_{jk} a_j(x)) \tag{23}$$

La Ecuación 22 corresponde a las ecuaciones que describen el primer momento del sistema, el cual corresponde a la media de la concentración de cada una de las especies que interactúan en la red reacciones. Por su parte, la Ecuación 23 permite obtener las ecuaciones que describen el segundo momento del sistema. Las ecuaciones del segundo momento corresponden al cruce de las especies en pares y representa la varianza de concentración de cada especie. El número de ecuaciones que se obtienen en el segundo momento obedece a una combinación con repetición de la forma $(n+1)!/2!\,(n-1)!$ donde n es el número de especies. En las Ecuaciones 22 y 23, v_j y $a_j(x)$ tienen el mismo significado que en la Ecuación 7. Los subíndices i y k representan una de las especies moleculares $\{S_1, \dots, S_N\}$.

Basado en la solución del sistema de ecuaciones obtenido anteriormente, se toman los momentos de la especie de interés. Sobre esta especie se realizará la inferencia de parámetros del sistema. Usualmente, la respuesta del sistema es representada por la especie molecular que corresponde a la expresión de una determinada proteína p. Partiendo de dicha especie, se denota la respuesta del sistema como m_y y M_y, los cuales corresponden al primer y segundo momento respectivamente. De este modo, se tiene que el momento m_y es igual a p^1. Este momento es equivalente a la respuesta promedio del sistema. Por su parte, M_y es el segundo momento "centrado" de la respuesta del sistema y es equivalente a la varianza teniendo en cuenta parámetros de ruido aditivo e_a y ruido multiplicativo e_b. Así, dichos momentos se expresan de la forma:

$$m_y = p^1 \tag{24}$$

$$M_y = (1 + e_b^2)var(p^1) + (e_a + e_b p^1)^2 + (p^1)^2 \tag{25}$$

Donde $var(p^1) = p^2 - (p^1)^2$. En estas ecuaciones el superíndice denota el orden de la variable. Posteriormente, es posible calcular la desviación estándar de la respuesta del sistema como $\sigma_y = \sqrt{M_y - (m_y)^2}$. La solución de la MC hasta un segundo orden parte del hecho de suponer que la respuesta del sistema se encuentra distribuida normalmente (González, 2014).

A partir de los valores de m_y y σ_y obtenidos de la respuesta del sistema es posible obtener el conjunto de parámetros que se ajusten a su comportamiento real. Para esto se toman estos valores, y se comparan con sus homólogos empíricos obtenidos mediante observaciones experimentales del sistema. En la herramienta estas observaciones corresponden a las observaciones simuladas mediante alguno de los algoritmos explicados anteriormente. Por lo tanto, se tiene que $\hat{m}_y$ y $\hat{\sigma}_y$ representan los valores de la media y la desviación estándar de las observaciones experimentales sobre la población.

De este modo, se comparan las observaciones con la aproximación basada en momentos a través de una función de costo que representa como se relacionan ambos conjuntos de datos. En este caso, se utiliza la medida de divergencia de Kullback-Leibler (KLD). Mediante esta medida, se calcula la relación de semejanza de la distribución de probabilidad $p = \{\hat{m}_y, \hat{\sigma}_y\}$ a la distribución modelo $q = \{m_y, \sigma_y\}$. Las distribuciones p y q contienen vectores que caracterizan la respuesta del sistema en cada instante de tiempo j hasta el final del experimento y/o simulación. La medida de la KLD es igual a cero cuando las distribuciones son idénticas e incrementa proporcionalmente en función de sus diferencias. Para la minimización de la KLD se utiliza un algoritmo de optimización en el cual se estiman los parámetros mediante la Ecuación 5. Donde el término $KLD(p||q)$ es equivalente a la siguiente expresión (González, 2014):

$$KLD(p||q) = \log\left(\frac{\sigma_y}{\hat{\sigma}_y}\right) + \frac{\hat{\sigma}_y{}^2 + (\hat{m}_y - m_y)^2}{2\sigma_y{}^2} - \frac{1}{2} \tag{26}$$

7.1.8 Inferencia de parámetros: Dos-Etapas

La inferencia de parámetros por Dos-Etapas trata de caracterizar una población en un enfoque de modelo de efectos mixtos. Este método consiste en obtener el estimado individual de parámetros ψ_i mediante el ajuste a cada individuo, para posteriormente, calcular directamente datos estadísticos de la población que

permitan determinar los parámetros que la caracterizan. Los parámetros individuales se obtienen aplicando la MLE a cada individuo, usando de forma similar las Ecuaciones 18 y 19. Sin embargo, se reducen las dimensiones de los datos observados a una sola dimensión. Esto debido a que solo se toma en cuenta un individuo observado a la vez. De este modo, la MLE de los parámetros $\theta_i = \{\psi_i, e_i\}$, dada la observación Y_i es la siguiente:

$$\widehat{\theta_i} = \arg\min(-\log(L(\theta_i, Y_i))) \tag{27}$$

$$-\log\big(L(\theta_i, Y_i)\big) = \sum_{j=1}^{T} \frac{1}{2}\left(\frac{Y_{ij} - f\big(t_j, u, x_0, \psi_i\big)}{h\big(f\big(t_j, u, x_0, \psi_i\big), e_i\big)}\right)^2 + \sum_{j=1}^{T} \log\left(h\big(f\big(t_j, u, x_0, \psi_i\big), e_i\big)\right) \tag{28}$$

Asumiendo que el modelo poblacional tiene una distribución log-normal, la media poblacional μ y la covarianza Ω de los parámetros se calculan de la siguiente forma:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \varphi_i \tag{29}$$

$$\Omega = \frac{1}{N-1}\sum_{i=1}^{N} (\varphi_i - \mu)(\varphi_i - \mu)^T \tag{30}$$

Donde $\varphi = \log(\psi)$. Igualmente, es posible determinar los parámetros de error, asumiendo que estos se encuentran regidos por un modelo de ruido constante a lo largo de la población. Así, los parámetros e_a y e_b se estiman como:

$$e_a = \frac{1}{N}\sum_{i=1}^{N} e_{a,i} \tag{31}$$

$$e_b = \frac{1}{N}\sum_{i=1}^{N} e_{b,i} \tag{32}$$

7.2 VERIFICACIÓN DE LAS UNIDADES

A continuación, se muestra la información que se obtuvo de la implementación de cada una de las unidades después de la búsqueda de errores y pruebas unitarias. Para la verificación de las unidades se utilizó como ejemplo el caso de estudio de expresión génica descrito en (Gonzáles, 2014). El proceso biológico descrito allí corresponde al proceso de la osmorregulación en levadura. Este proceso conlleva

una de las rutas de señalización de proteína quinasas activadas por mitógenos (MAPK), conocida como la ruta del glicerol hiperosmótico (HOG). Dicha ruta es activada en respuesta a un estímulo de tipo shock hiperosmótico, con el fin de mantener la homeostasis de agua en el ambiente intracelular. La activación del HOG desencadena una serie de respuestas en el citoplasma, tal como la modificación de expresión de una gran cantidad de genes mediante la acción de factores de transcripción. De este modo, se genera la acumulación de glicerol, que conduce al restablecimiento del equilibrio osmótico al interior de la célula. Uno de los genes inducidos es el gen STL1, el cual es modificado para que codifique la expresión de una proteína fluorescente, conocida como yECitrine; permitiendo observar la respuesta del sistema biológico mediante la medición de los niveles de fluorescencia de cada célula (Gonzáles, 2014).

El proceso de osmorregulación en levadura puede ser descrito en forma de una red de reacciones químicas, en la que se representa la forma en que las especies moleculares reaccionan entre sí. Mediante esta representación es posible observar de forma clara la evolución de cada una de las especies; es decir, su producción y degradación a lo largo del proceso. Para la construcción del modelo estructural o base del proceso se toma la siguiente red de reacciones:

$$\emptyset \xrightarrow{c1*u(t)} mRNA \quad \text{(R1)}$$

$$mRNA \xrightarrow{c2} \emptyset \quad \text{(R2)}$$

$$mRNA \xrightarrow{c3} mRNA + Proteína \quad \text{(R3)}$$

$$Proteína \xrightarrow{c4} \emptyset \quad \text{(R4)}$$

Las reacciones R1 y R2 representan la producción y degradación de *mRNA*. Como se puede observar R1 depende de un estímulo de entrada $u(t)$. Por su parte, la reacción R3 representa cómo a partir de *mRNA* se obtiene este mismo más la expresión de *Proteína*. Esta última, a la vez se degrada en R4. Desde esta red de reacciones se obtiene la información necesaria para realizar el modelado de un sistema biológico en la herramienta.

7.2.1 Definición del Sistema Biológico

De la red de reacciones establecida se tiene que, las especies moleculares son *mRNA* y *Proteína.* Los parámetros cinéticos corresponden a $c1 = 4$, $c2 = 0.01$, $c3 = 1$, y $c4 = 0.006$; donde $c1$ y $c3$ representan producción, y $c2$ y $c4$, degradación de las especies respectivamente. Asimismo, se obtiene las matrices R y P; las cuales representan los coeficientes estequiométricos de las especies moleculares cuando

actúan como reactivos o productos de una reacción. Por reactivos se entienden las especies que se encuentran a la izquierda de la red de reacciones; mientras que, por productos, las especies que se encuentran a la derecha. Los coeficientes estequiométricos toman el valor de 1, si una determinada especie participa como reactivo o producto en cada una de las correspondientes reacciones.

$$R = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Como resultado de la información anterior, se obtuvo la matriz estequiométrica general $V = P - R$ y el sistema de ecuaciones diferenciales que se observan en las Figura 10 respectivamente.

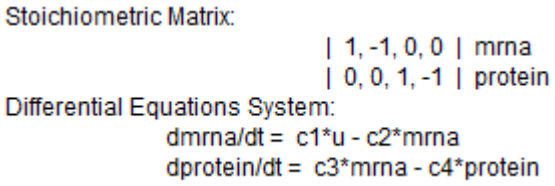

Figura 10. **Matriz estequiométrica y sistema de ecuaciones diferenciales.**

7.2.2 Simulación de la Población Celular

Posteriormente, se definen las propiedades del sistema relacionadas con la etapa de simulación. En primer lugar, se define la duración del experimento; la cual en este caso tiene una duración de 100 minutos. También, se definen las concentraciones iniciales de las especies como: *mRNA* = 0 y *Proteína* = 0. En total, se simuló una población celular de 10000 células.

Una vez definidas las propiedades del sistema, se establece qué tipo de entrada estimula el sistema. Siguiendo, lo planteado en el ejemplo se aplica una entrada de tipo modelo y que sigue la ecuación diferencial presentada en la Ecuación 10. La entrada se denomina como $HOG(t)$ y tiene los parámetros de producción y degradación $k = 0.3968$ y $g = 0.9225$. Como se mencionó anteriormente, el perfil de entrada sigue su propio modelo; sin embargo, se define de la misma forma que la entrada de tipo pulso. Luego, se establecen el pulso de shock osmótico en el minuto 1; y se aplica durante 4 minutos. De este modo, se obtiene el perfil del estímulo de entrada que se muestra en la Figura 11; donde se puede observar la evolución en el tiempo del estado del pulso de shock osmótico y la expresión del gen generado por el *HOG.*

***Figura 11*. Perfil del estímulo de entrada simulado.**

A continuación, se muestran las respuestas obtenidas desde los algoritmos implementados para simular las diferentes fuentes de variabilidad. La respuesta observada corresponde a la expresión de la especie *Proteína*. En primer lugar, se observa la respuesta simulada sin tener en cuenta ningún tipo de variabilidad, lo cual conlleva a que sea de naturaleza determinista y siempre se comporte igual, tal como se muestra en la Figura 12. Si se simulara una población entera de este modo, todas las células obtendrían la misma expresión debido a que comparten el mismo conjunto de parámetros. Por lo tanto, para obtener respuestas individuales diferentes, se aplica sobre el sistema una de las fuentes de variabilidad detalladas anteriormente. En primer lugar, se puede aplicar un modelo de ruido y definir los parámetros de ruido de medición. En este caso se estableció un ruido aditivo $e_a = 80$, y ruido multiplicativo $e_b = 0.05$. La respuesta individual de la población obtenida después de añadir el ruido de medición se puede observar en la Figura 13 (A). Lo anterior permite que en la Figura 13 (B), se pueda observar un comportamiento a nivel poblacional de la respuesta del sistema. Donde la línea azul oscuro representa el equivalente a la mediana de la respuesta; mientras que la sombra azul representa al 95 % de la población total simulada.

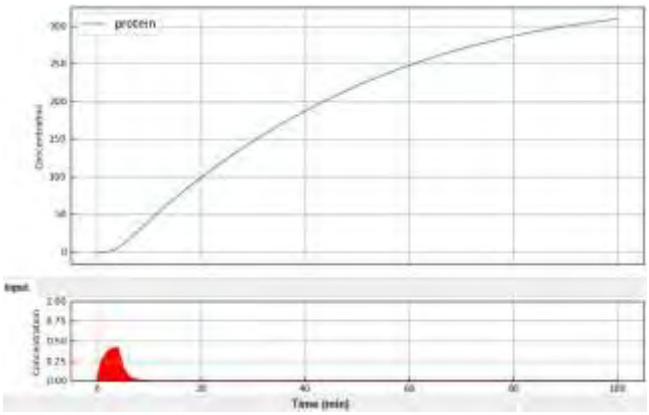

***Figura 12.* Simulación determinista del proceso biológico.**

Por su parte, en las Figuras 14 (A) y 14 (B) se observa la respuesta individual y poblacional obtenida a partir de la simulación de estocasticidad intrínseca. Para este tipo de variabilidad no es necesario definir ningún parámetro adicional. En este caso se observa como la expresión de cada célula se diferencia de las demás. Esto se debe a que la aproximación realizada mediante el algoritmo de Gillespie toma de forma aleatoria cada una de las reacciones en diferentes instantes de tiempo.

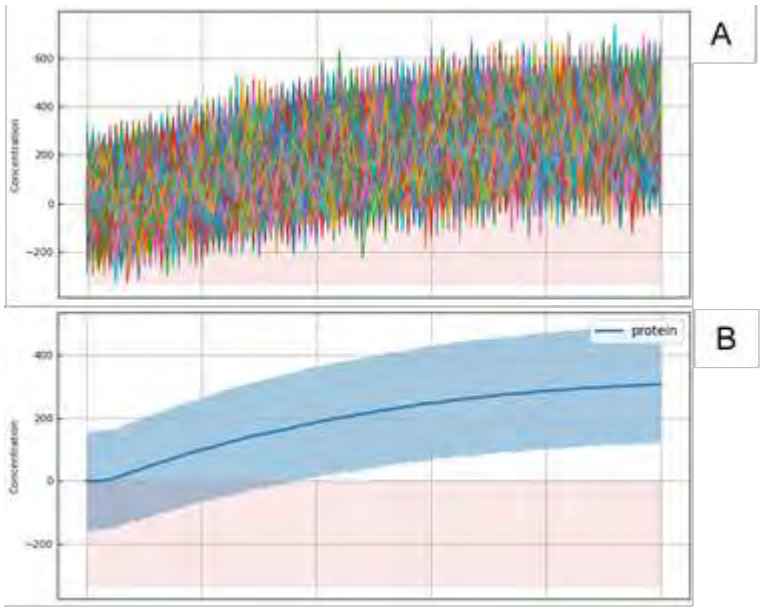

Figura 13. **Simulación de ruido de medición. A) Respuestas individuales. B) Respuesta poblacional.**

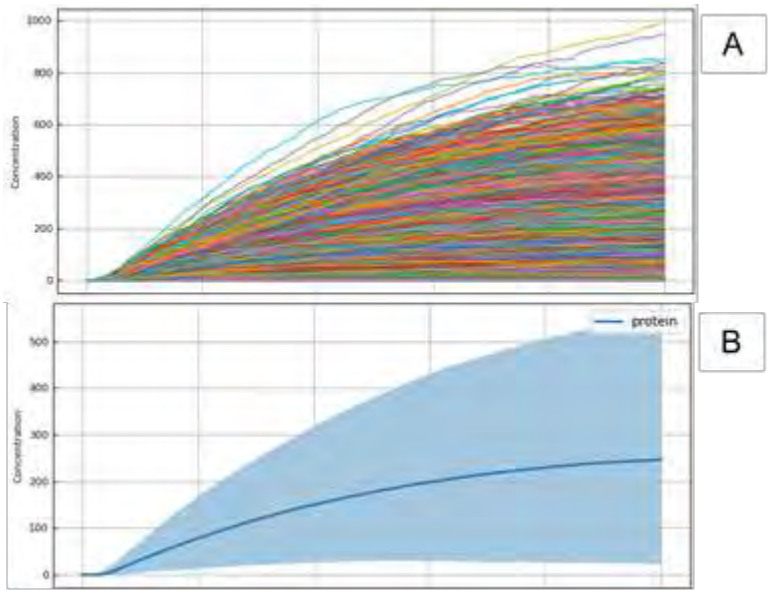

Figura 14. **Simulación de estocasticidad intrínseca. A) Respuestas individuales. B) Respuesta poblacional.**

Por último, para la variabilidad extrínseca se definió la matriz de covarianza Ω, la cual es una matriz identidad con sus valores con valores de 0.05 en su diagonal. Esta matriz es de naturaleza logarítmica y debe ser simétrica semidefinida positiva.

De este modo, con Ω y el logaritmo de los parámetros cinéticos se obtuvo la distribución normal de los parámetros. La matriz de covarianza Ω se estableció como una matriz diagonal con valores de 0.05. En las Figura 15 (A) se puede observar como la distribución de parámetros permite obtener una distribución de la respuesta individual del sistema. Por su parte, la Figura 15 (B) muestra la respuesta poblacional obtenida a partir de variabilidad extrínseca.

$$\Omega = \begin{bmatrix} 0.050 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.050 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.050 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.050 \end{bmatrix}$$

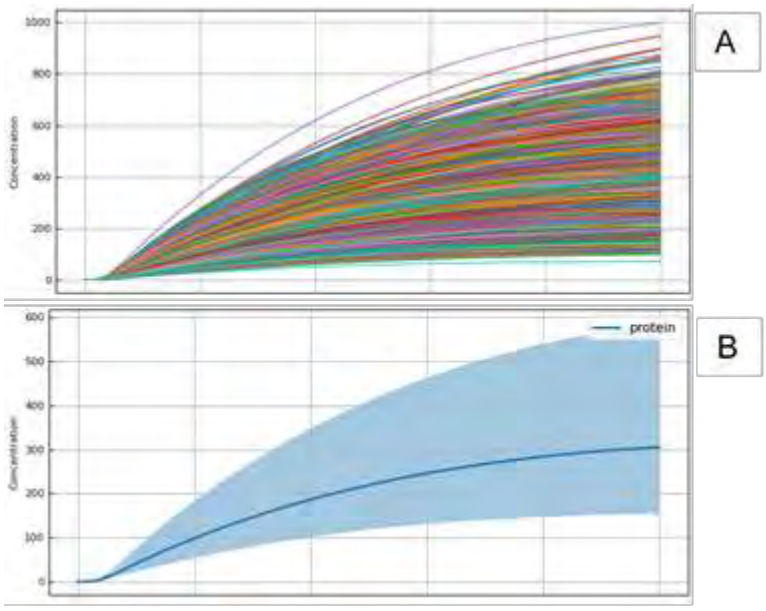

Figura 15. **Simulación de variabilidad extrínseca. A) Respuestas individuales. B) Respuesta poblacional.**

Desde la Figura 13 a la 15 se mostró la respuesta de los algoritmos implementados para simular los diferentes tipos de variabilidad. Dichas figuras muestran la simulación individual de cada tipo; sin embargo, mediante la herramienta es posible conjugarlas para obtener una respuesta que influenciada por diferentes fuentes de variabilidad. No obstante, esto puede dificultar la posterior inferencia de parámetros sobre las observaciones simuladas.

7.2.3 Inferencia de Parámetros

Para la inferencia de parámetros se implementaron varios tipos de modelos que caracterizan la respuesta del sistema dependiendo del tipo de variabilidad que lo afecta; es decir, que la eficiencia de cada método para estimar los parámetros se encuentra influenciada por la fuente de variabilidad que se simuló en la etapa previa. Los métodos o modelos implementados son:

- *Ajuste de curva media*, este método es el más simple de los implementados. Como tal no infiere la variabilidad del sistema; sin embargo, sirve como apoyo para encontrar los parámetros cinéticos o puntos de partida para los valores iniciales aplicados a los otros modelos. El ajuste de curva media usa mínimos cuadrados de forma no lineal para ajustar una función a un conjunto de datos. Adicionalmente, este método puede ser considerado como una alternativa al modelo de Célula promedio.

- *Célula promedio,* este modelo estima mejor la variabilidad proveniente del ruido de medición. La estimación de parámetros se obtiene mediante la minimización de la MLE.

- *Basado en momentos,* captura la variabilidad de sistemas que presentan mayormente naturaleza estocástica intrínseca. La inferencia mediante este modelo parte de calcular los momentos del sistema hasta el segundo orden para luego estimar los parámetros mediante la minimización de la KLD.

- *Dos-Etapas,* estima de forma individual cada individuo de la población para capturar la variabilidad extrínseca. A partir de dichas estimaciones, calcula los parámetros poblacionales que caracteriza el sistema.

A continuación, se muestran los resultados obtenidos a partir de simulaciones de sistemas. Las observaciones obtenidas corresponden a sistemas con variabilidad aplicada de la forma: ruido de medición, estocasticidad intrínseca, estocasticidad intrínseca-ruido de medición, variabilidad extrínseca, y variabilidad extrínseca-ruido de medición. El sistema que solo posee ruido de medición se utilizó para realizar la inferencia a través de *Ajuste de Curva Media* y *Célula Promedio*. Por su parte, con los sistemas con estocasticidad intrínseca y estocasticidad intrínseca-ruido de medición se estimaron utilizando el modelo *Basado en momentos*. Finalmente, se utilizó el modelo de Dos-Etapas para estimar los sistemas con variabilidad extrínseca y variabilidad extrínseca-ruido de medición. Dichos sistemas fueron simulados a partir de una matriz de covarianza diagonal con valores de 0.05. En

cuanto a los parámetros de ruido, todos los sistemas fueron simulados con valores de 80.0 y 0.05 correspondientes a ruido aditivo y ruido multiplicativo respectivamente. Para todos los ejemplos que se muestran a continuación se tomó como parámetros cinéticos iniciales el vector $beta0 = [3.0, 0.02, 0.9, 0.01]$ y parámetros de ruido iniciales $e_a = 30.0$ y $e_b = 0.1$, para los sistemas con ruido; y $e_a = 0.1$ y $e_b = 0.1$, para los sistemas sin ruido. Para la inferencia de los parámetros se utilizó un algoritmo de optimización downhill simplex, se establecieron un total 2000 iteraciones máximas a ejecutar para cada estimación con un valor de tolerancia de 0.0001.

En las Figuras 16 (A) y 16 (B) se muestran las estimaciones obtenidas mediante el método de Ajuste *de Curva Media* y *Célula Promedio*. Como se puede observar, la estimación (líneas naranjas) en ambos casos se ajustaron perfectamente a los datos observados (líneas azules). Sin embargo, tal como se muestra en la Tabla 3, los parámetros inferidos $c1$ y $c3$ difieren a los parámetros reales en el modelo de *Célula Promedio*, probablemente debido a que el algoritmo de optimización se detuvo en un mínimo local. Otro factor que influye en esta divergencia se encuentra relacionada con el hecho de que son parámetros multiplicativos, lo que ocasionan que estos puedan ser compensados. Por su parte, cabe resaltar que mediante el modelo de *Célula Promedio* se pudo inferir exactamente los parámetros de ruido de medición.

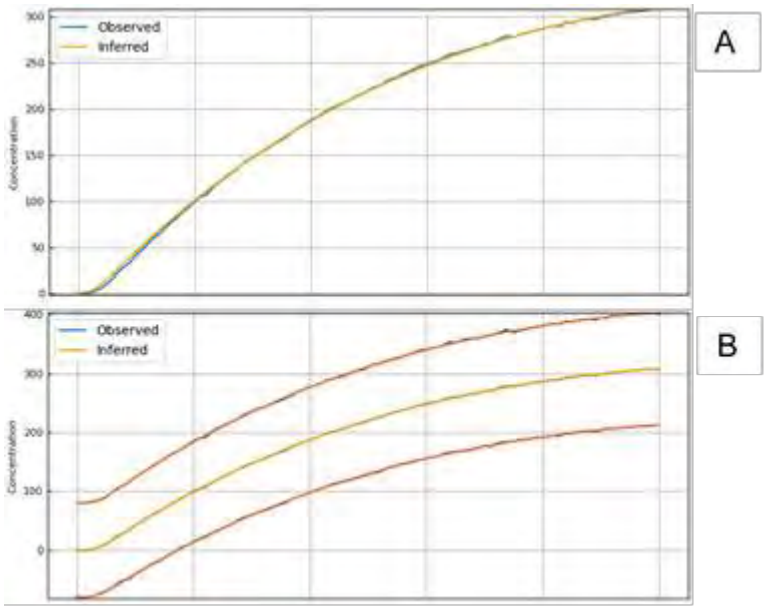

Figura 16*. Estimación de ruido de medición. A) *Ajuste de Curva Media.* B) *Célula Promedio.

En los modelos estocásticos debido a la aleatoriedad de su respuesta; es necesario aproximar el modelo del sistema de forma determinística. Para lo cual, se realizó dicha aproximación mediante el cálculo de los momentos estadísticos del sistema biológico. Para la inferencia basada en momentos se tomó el primer y segundo momento de la especie de salida. Estos corresponden a la media y la varianza respectivamente. En la Figura 17, se muestra el sistema de ecuaciones diferenciales a partir del cual se realizó la minimización de KLD para la inferencia de parámetros. Mediante este modelo se obtuvieron los parámetros inferidos que se muestran en la Tabla 3. De igual modo, este método, ajustó las observaciones exitosamente y obtuvo valores inferidos similares a los reales; a excepción del parámetro $c1$. Una razón para esto puede ser que en el espacio de parámetros en el que se encuentran los valores reales, existen varios óptimos locales en los cuales converge el algoritmo de optimización. El ajuste de los perfiles obtenidos se muestra en las Figuras 18 (A) y 18 (B). Al comparar los dos sistemas con estocasticidad intrínseca se puede resaltar que, aunque solo uno de ellos tenía ruido de medición, ambos convergieron a un conjunto de parámetros cinéticos similar. Por su parte, el sistema que solo poseía estocasticidad intrínseca aproximó los parámetros de ruido a valores cercanos a cero.

```
Differential Equations of Moments:
                dmrna/dt =  c1*u - c2*mrna
                dprotein/dt =  c3*mrna - c4*protein
                dmrna**2/dt =  2*c1*mrna*u + c1*u - 2*c2*mrna**2 +
c2*mrna
                dmrna*protein/dt =  c1*protein*u - c2*mrna*protein +
c3*mrna**2 - c4*mrna*protein
                dprotein**2/dt =  2*c3*mrna*protein + c3*mrna -
2*c4*protein**2 + c4*protein
```

Figura 17. **Sistema de ecuaciones diferenciales de los momentos del sistema.**

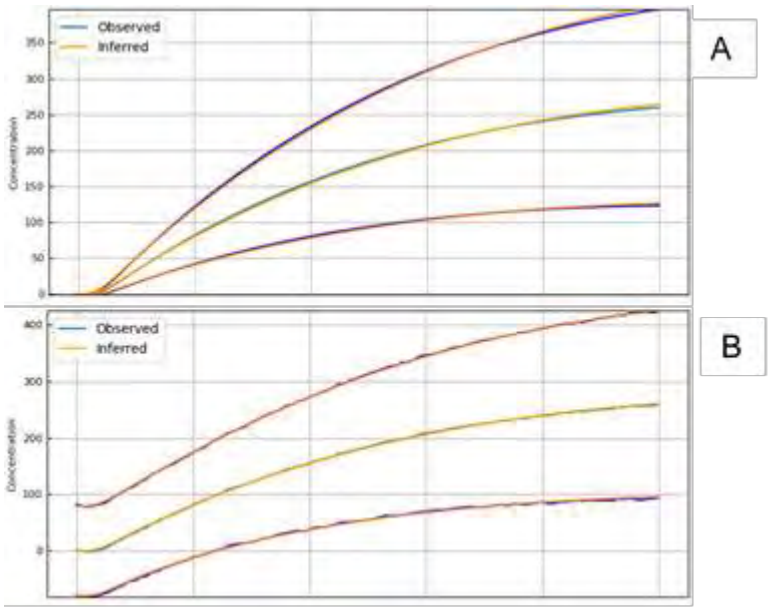

Figura 18. **Estimación mediante modelo *Basado en Momentos.* A) Estocasticidad intrínseca. B) Estocasticidad intrínseca-ruido.**

Particularmente, para el método de *Dos-Etapas* se definió una subpoblación de 100 células para ser estimadas individualmente. Debido a que a nivel individual el número de iteraciones de estimación es menor, se estableció un total de 300 iteraciones para cada célula. Al final de la estimación se obtiene la comparación entre la población real (líneas azules), la subpoblación observada (líneas verdes) y el perfil inferido (líneas naranjas), tal como se muestra en la Figuras 19 (A) y 19 (B). Como se puede observar en la Tabla 3, este método obtuvo los parámetros inferidos más cercados a los valores reales, tanto para los cinéticos como para los de ruido de medición. Sin embargo, al comparar las observaciones con los perfiles inferidos se puede percibir que hay cierta discrepancia especialmente a los que conciernen con la varianza de la expresión poblacional. Esto puede deberse a los valores obtenidos en la matriz de covarianza inferida Ω_{inf2}, los cuales representan una mayor varianza en los parámetros c_2 y c_4 con respecto a los demás parámetros. Por otro lado, la estimación mediante el modelo de *Dos-Etapas* obtuvo un mejor desempeño con el sistema sin ruido de medición. Esto puede deberse a el efecto que conllevan los parámetros de ruido y su influencia sobre la expresión general del sistema.

$$\Omega_{inf1} = \begin{bmatrix} 0.214 & -0.048 & -0.165 & 0.078 \\ -0.048 & 0.450 & 0.073 & -0.198 \\ -0.165 & 0.073 & 0.215 & -0.095 \\ 0.078 & -0.198 & -0.095 & 0.213 \end{bmatrix}$$

$$\Omega_{inf2} = \begin{bmatrix} 0.335 & 0.203 & -0.264 & -0.178 \\ 0.203 & 2.037 & 0.063 & -0.513 \\ -0.264 & 0.063 & 0.340 & 0.125 \\ -0.178 & -0.513 & 0.125 & 1.174 \end{bmatrix}$$

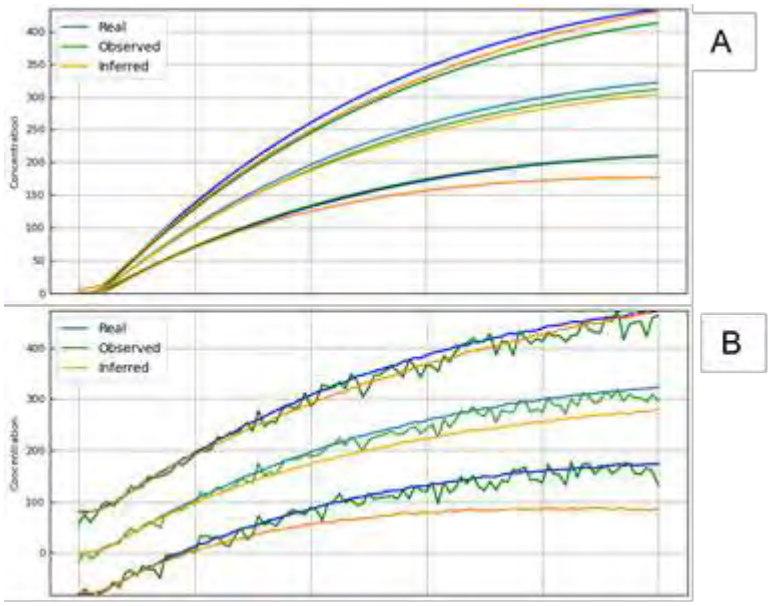

Figura 19. **Estimación mediante modelo *Dos Etapas.* A) Variabilidad extrínseca. B) Variabilidad extrínseca-ruido.**

En la Tabla 3, se resaltan los parámetros que más se aproximaron a sus valores reales. En general, los modelos permitieron converger a un conjunto de parámetros que se ajustara a las observaciones. Por su parte, en los sistemas sin ruido se esperaba que los valores $\boldsymbol{e_a}$ y $\boldsymbol{e_b}$ obtuvieran valores cercanos a cero; representando la ausencia de ruido. Otro aspecto para resaltar es el comportamiento similar de los modelos ante un sistema con ruido y uno sin ruido. Esto se debe a que se tuvo precaución con la sensibilidad de los modelos ante los parámetros iniciales. Por ende, se tuvo en cuenta el hecho de que un sistema tuviera ruido para definir el punto de partida del algoritmo de optimización.

Tabla 3.

Resultados de estimación de parámetros.

			Modelo					
			Célula Promedio	Basado en Momentos	Dos-Etapas		Basado en Momentos	Dos-Etapas
			Tipo de variabilidad					
			Ruido	*Ruido-Intrínseca*	*Ruido-Extrínseca*		*Intrínseca*	*Extrínseca*
Parámetros	**Valores reales**	**Valores iniciales**	**Valores inferidos**			**Valores iniciales**	**Valores inferidos**	
c_1	4.000	3.000	0.524	2.935	3.638	3.000	2.713	3.597
c_2	0.010	0.020	0.010	0.008	0.009	0.020	0.006	0.005
c_3	1.000	0.900	7.641	1.120	1.144	0.900	1.175	1.069
c_4	0.006	0.010	0.005	0.006	0.005	0.010	0.008	0.009
e_a	80.000	30.000	80.050	79.576	79.242	0.100	0.010	0.129
e_b	0.050	0.100	0.049	0.069	0.064	0.100	0.028	0.002

En conjunto con la inferencia de parámetros, es posible realizar un análisis de estos. De tal forma, que se visualice la evolución de cada uno de los parámetros a lo largo de las iteraciones llevadas a cabo por el algoritmo de optimización. En la Figura 20 se muestra un ejemplo del análisis de parámetros que se obtuvo de la inferencia realizada sobre la población usada para inferir los parámetros mediante el modelo de *Célula Promedio*. Mediante este gráfico es posible percibir fácilmente hacia donde tiende el valor óptimo de cada uno de los parámetros. Además, permite observar, como el hecho de que usualmente los parámetros son términos multiplicados, ocasiona que se traten de compensar el uno al otro; llevando de este modo, a que el algoritmo de optimización se estanque en un óptimo local. Otro aspecto para analizar concierne a que tan distante pueden estar cada uno de los parámetros iniciales de su valor real; para que el algoritmo no pueda minimizar efectivamente la función de costo del sistema, y determinar qué tan sensible es a dicho distanciamiento.

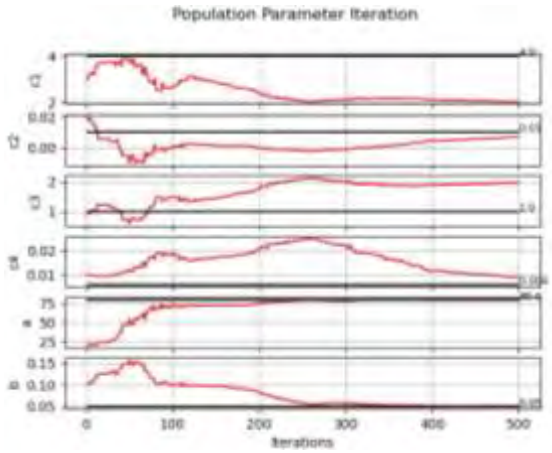

Figura 20. **Análisis de parámetros.**

7.3 INTEGRACIÓN DE LAS UNIDADES DE LA HERRAMIENTA

Las unidades implementadas se agruparon en una interfaz gráfica de usuario (GUI), la cual fue diseñada mediante el software *Qt Designer* y programada usando la librería *PyQt5*. Esta librería permitió enlazar cada uno de los objetos en la GUI con el código *Python* que lleva a cabo las funciones propuestas para la herramienta. La Figura 21 muestra la interfaz obtenida como resultado del desarrollo de la herramienta. La GUI cuenta con cuatro elementos principales: 1) Ventana de ingreso de información para definir las características de un sistema biológico. 2) Ventana de propiedades del sistema que muestra los datos registrados y las ecuaciones o variables calculadas a partir de la información inicial. 3) Comandos para ejecutar cada una de las secciones de la herramienta. 4) Graficación de la respuesta individual, respuesta poblacional y entrada del sistema, y análisis de parámetros. 5) Barra de tareas que permite guardar o abrir sistemas simulados. Las diferentes gráficas contenidas en el cuarto elemento muestran el cambio de concentración de las especies en función del tiempo en minutos.

La ventana de ingreso de información contiene tres secciones definidas como *definir* (Def), *simular* (Sim) e *inferir* (Inf). La sección *definir* se puede observar en la Figura 21. En esta sección se ingresa la información necesaria para definir las especies moleculares y los parámetros cinéticos de las reacciones que se van a modelar. De igual forma, se define la reacción con la que interactúa la entrada del sistema, y las matrices de reactivos y productos para definir como es el cambio de estado de las especies moleculares involucradas.

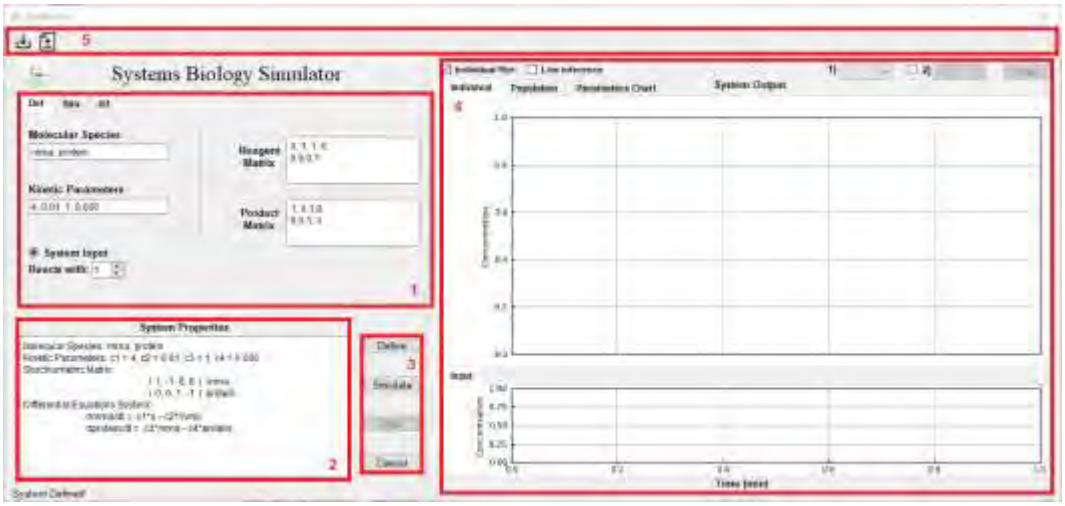

Figura 21. **Elementos principales de la GUI de la herramienta. Componentes de la sección "definir".**

Por su parte, en la sección *simular* se programa la simulación del sistema definido previamente. En un principio, se establece la duración del experimento, la concentración inicial de las especies moleculares, y el número de células que componen el sistema. Luego, se establece la configuración correspondiente al estímulo de entrada del sistema; para lo cual se define el tipo de entrada (pulso o modelo), y el inicio y la duración de los pulsos. Por último, es posible seleccionar las fuentes de variabilidad que se desean añadir a la simulación del sistema y la respuesta de una determinada especie a visualizar al final de la simulación. Los componentes de esta sección se pueden observar en la Figura 22.

Figura 22. **Componentes de la sección "simular".**

Finalmente, en la Figura 23 se muestra la sección *inferir.* En esta sección se definen los parámetros cinéticos iniciales, los parámetros de ruido, el número de iteraciones, el valor de tolerancia y el tipo de modelo que se va a usar para estimar y/o caracterizar el sistema biológico observado. Para el caso del modelo de *Dos-Etapas,* es posible establecer el número de células usadas para la estimación. El proceso de inferencia se visualiza paralelamente en la gráfica de la respuesta del sistema; de tal modo que, es posible observar el ajuste de la inferencia a la respuesta simulada previamente.

Mediante los comandos que se encuentran en la parte superior derecha de la gráfica de respuesta del sistema es posible manipular y seleccionar la especie a visualizar. Adicionalmente, permite visualizar todas o ninguna de las especies a la vez en una misma gráfica o graficar una especie secundaria en un gráfico adicional.

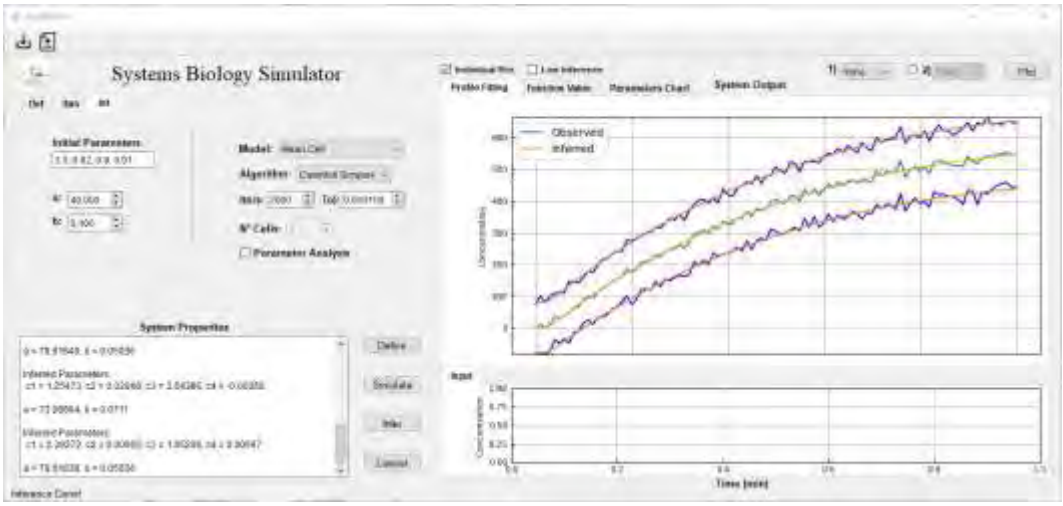

Figura 23. **Componentes de la sección "inferir".**

8. VALIDACIÓN DE LA HERRAMIENTA

Posterior a la implementación de la herramienta, se realizó la documentación necesaria para su uso, de tal forma que se detalla de forma concisa la forma de inicializar la herramienta y los aspectos relevantes de esta. Por otra parte, se desarrolló una experiencia educativa en la que se explica paso a paso como usar la herramienta a partir de casos de estudio propuestos. De este modo, se evaluaron a los estudiantes involucrados en la experiencia con el objetivo de validar la herramienta como parte del proceso de aprendizaje de sistemas biológicos.

8.1 DOCUMENTACIÓN TÉCNICA

La documentación técnica de la herramienta es de pertinencia para cualquier persona que se encuentre interesada en aprender o que tenga conocimientos básicos en el área de la biología de sistemas, específicamente en los aspectos que conciernen a la simulación y modelado de procesos biológicos en poblaciones celulares. Este documento se muestra en el Anexo A y presenta la forma adecuada de ejecutar la aplicación e interactuar con ella. Para ello, se realiza una descripción de cada uno de los elementos en la interfaz gráfica, el significado de cada una de las variables de entrada y las opciones de simulación, modelado y visualización disponibles.

8.2 PRUEBAS DE USABILIDAD

Para la validación de la usabilidad de la herramienta se diseñó una experiencia educativa. En el Anexo B se muestra el documento que tiene como fin utilizar la herramienta en varios casos de estudios relacionados con expresión génica en poblaciones celulares. De igual forma, en el documento se incluyeron diferentes sistemas biológicos relacionados con otros procesos como reacciones enzimáticas, crecimiento de diferentes poblaciones, y procesos metabólicos. A partir de estos casos se pretende probar la usabilidad de la herramienta en diferentes tipos de sistemas biológicos.

El desarrollo de la experiencia educativa consistió en un seminario-taller que tuvo lugar de forma online. Durante la experiencia participaron un total de 16 estudiantes de ingeniería biomédica entre 7 y 9 semestre de la Universidad Autónoma de Occidente. La actividad inició con una presentación e introducción a la biología computacional, a la biología de sistemas y al proceso de modelado de sistemas biológicos en poblaciones celulares. Los conceptos presentados, luego fueron relacionados con las características de la herramienta, y se propuso un taller

individual en el cual los estudiantes debía abordar desde la herramienta un caso de estudio. Previo y durante al taller cada estudiante tuvo acceso a la herramienta e igualmente pudo interactuar con ella desde sus computadoras. Al finalizar la actividad, cada estudiante realizó una encuesta y un quiz con los cuales se buscó; primero, evaluar la usabilidad de la herramienta y segundo, evaluar la apropiación de conocimientos de los estudiantes a partir de su interacción con la herramienta y la observado en la presentación.

La encuesta desarrollada se muestra en el anexo D, en ella se preguntó información básica del estudiante y se pidió calificar cada una de las características principales de la herramienta en relación con su relevancia en el proceso de aprendizaje que llevaron a cabo durante la experiencia. Por su parte, el quiz se muestra en el anexo E. Este quiz constó de 5 preguntas abiertas que pretendían hacer reflexionar a los estudiantes sobre los conceptos aprendidos.

8.2.1 Resultados de la encuesta para evaluar la usabilidad de la herramienta

La Figura 24 muestra la evaluación de algunos aspectos generales de la herramienta. Cada aspecto fue evaluado acorde a una escala de 1 a 5, donde 5 es la mejor calificación y 1 la peor. De forma general, se puede decir que la herramienta tuvo una buena aceptación por parte de los estudiantes. Los aspectos para destacar, según la calificación de los estudiantes, son: confiabilidad, claridad de las secciones, relevancia para el aprendizaje y documentación. Por otra parte, aunque tuvieron una calificación alta; la estética de la interfaz de usuario, la facilidad de uso y la protección ante errores del usuario son aspectos que se pueden mejorar en la herramienta.

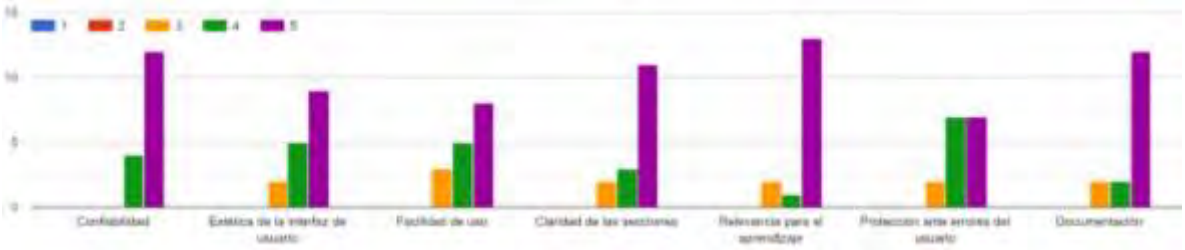

Figura 24. **Evaluación de aspectos generales de la herramienta.**

Posteriormente, se evaluó de forma específica algunas de las funcionalidades principales de la herramienta. La calificación se denota de acuerdo con la siguiente escala: 1 = Nada pertinente, 2 = Poco pertinente, 3 = Más o menos pertinente, 4 = Muy pertinente, 5 = Definitivamente pertinente.

En primera instancia, las pestañas de ingreso de información tuvieron una calificación media. Pues, aunque los estudiantes comentaron que era una funcionalidad útil e interesante para poder variar las variables y condiciones iniciales de diferentes sistemas, uno de ellos comentó que inicialmente tuvo complicaciones con ellas, pero igualmente logro ingresar los datos requeridos.

Por su parte, la definición del sistema tuvo una calificación alta. De forma general, los estudiantes expresaron la importancia de esta funcionalidad para comprender el funcionamiento matemático que puede tener un sistema, puesto que permite la adquisición fácil, rápida y confiable de las ecuaciones diferenciales y la matriz estequiométrica.

La funcionalidad calificada con mayor pertinencia fue la simulación de la población. Alrededor del 80 % de los estudiantes la calificaron como definitivamente pertinente. En cuanto a esta funcionalidad, se expresó su pertinencia en cuanto a que permite observar el comportamiento de una célula al ser estimulada.

Finalmente, la inferencia de parámetros también obtuvo una calificación alta de pertinencia. Esta funcionalidad fue percibida como muy útil, ya que permite obtener los parámetros que posibilitan aproximar un conjunto de observaciones a un determinado modelo.

A partir de la actividad los estudiantes, en general, concluyeron y resaltaron la utilidad de la herramienta para dar una introducción a la investigación biomédica enfocada en la simulación de sistemas biológicos y análisis de sus comportamientos en función de los parámetros y las fuentes de variabilidad que los afectan. Adicionalmente, los estudiantes encontraron interesante el tema abordado, y recalcaron lo eficaz que fue su proceso de aprendizaje mediante el uso de una herramienta. De este modo, se resaltó la importancia de la creación de nuevos softwares en diferentes áreas de la ciencia que sirvan como soporte a nivel profesional y como herramientas de aprendizaje.

Por otra parte, aunque la herramienta fue percibida como muy interactiva y fácil de usar, como recomendación los estudiantes concluyeron en que se podría mejorar la documentación de uso de la herramienta y en lo posible incluirla dentro del software. Con relación a la última recomendación, se expresó que sería ideal entregar información rápida sobre la interfaz o las variables en forma de iconos de interrogación (?), evitando la necesidad de buscar en la documentación. En cuanto a la interfaz, se sugirió que esta podría mejorarse, haciéndola más llamativa y agradable a la vista. Además, se recomendó añadir la opción de maximizar la GUI para que sea pueda observar mejor las gráficas y los cuadros de texto sobre la

interfaz. Adicionalmente, se recomendó mejorar la forma en la que se presentan los resultados, y ser más específico al momento de mostrar los errores. Por último, se sugirió poner el nombre completo de las pestañas de ingreso de información para mayor claridad y comodidad, y la creación de un video explicando el proceso de instalación de la herramienta.

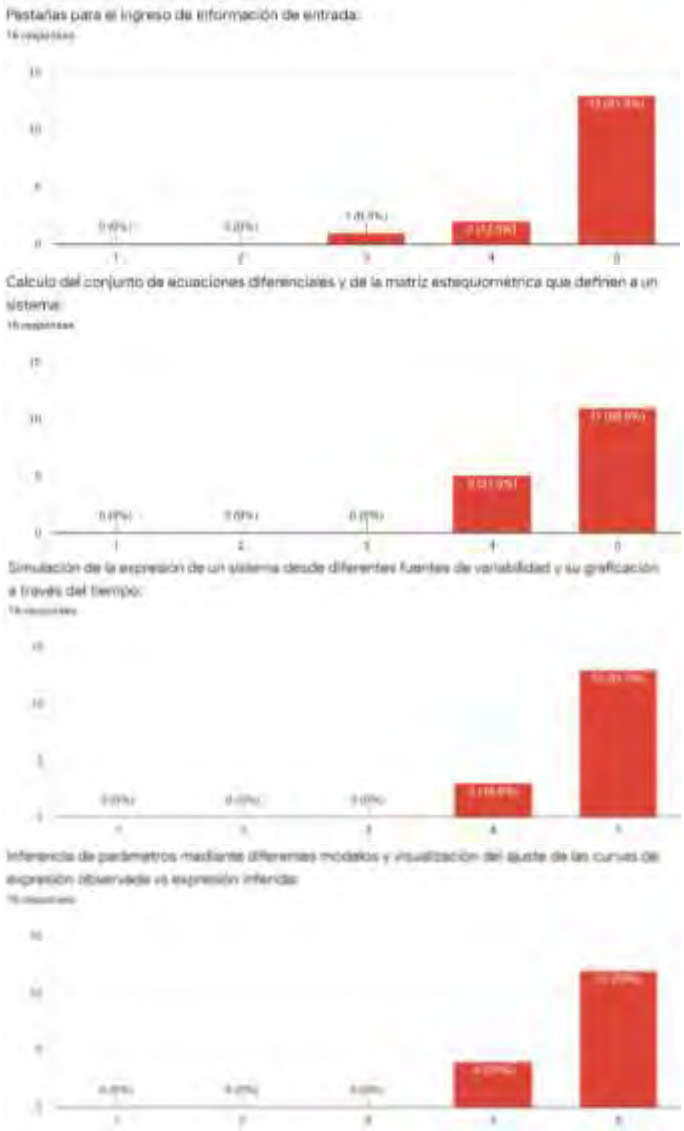

Figura 25. **Evaluación de: A) Pestañas de ingreso de información. B) Definición del sistema. C) Simulación de la población. D) Inferencia de parámetros.**

8.2.2 Resultados del quiz para evaluar la apropiación de conocimientos

El quiz fue usado para evaluar la apropiación de conocimientos durante la experiencia educativa. En la Figura 26 se muestran los resultados por pregunta, como se puede observar la pregunta 3, relacionada con el modelo de *Célula Promedio,* fue la pregunta en la cual un mayor porcentaje de estudiantes se equivocaron, siendo alrededor de 62.5% de la población.

El 75% de estudiantes aprobó el quiz y a nivel general el grupo tuvo un promedio de 3.43, por ende, se puede decir que la herramienta fue relevante en el proceso de aprendizaje sobre el tema. Para mejorar este porcentaje, se podrían realizar las mejoras sugeridas por los estudiantes en la encuesta, y analizar las causas por las cuales la pregunta 3 tuvo un bajo índice de aprobación.

Las preguntas realizadas en el quiz fueron:

1. La experiencia educativa se centró en un área de la biología computacional, ¿Cuál fue y en qué consiste?

2. Explique el "dogma" central de la biología molecular.

3. ¿Cómo se genera la variabilidad en una población celular mediante el modelo de célula promedio?

4. Desde un punto de vista biológico, ¿Cuáles son las causas de la variabilidad intrínseca?

5. ¿Cuáles son las fuentes de variabilidad extrínseca, y cómo se modelan con efectos mixtos?

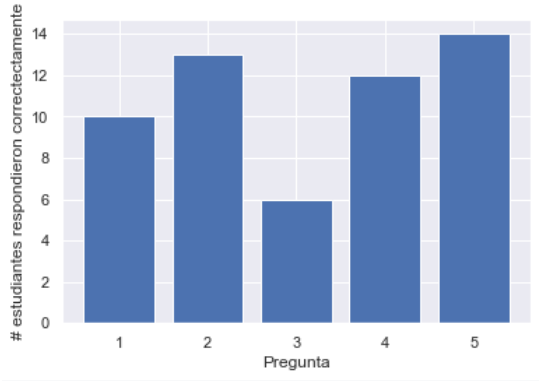

Figura 26. **Número de respuestas correctas por pregunta.**

9. CONCLUSIONES

El desarrollo de la herramienta estuvo regido por la información extraída de la literatura sobre el modelado de procesos biológicos, específicamente sobre procesos de expresión génica en poblaciones celulares. Mediante la herramienta es posible emular el paradigma típico usado en el área de la biología de sistemas. De este modo, la herramienta favorece el aprendizaje sobre un proceso biológico ya que permite que el usuario defina las partes que lo componen de tal forma que entienda la interacción de cada una de las moléculas que lo integran y las reacciones que se llevan a cabo. Esto con el fin de definir un sistema de ecuaciones diferenciales que permita representar la respuesta del sistema biológico. A partir de dichas ecuaciones la herramienta puede representar gráficamente simulaciones de la expresión de un determinado proceso, crear poblaciones celulares, y exhibir las diferentes fuentes de variabilidad que lo pueden afectar de forma determinista o estocásticamente. De igual manera, la herramienta logra de forma interactiva representar el proceso de modelado desde diferentes modelos, a partir de los cuales se obtienen los parámetros que caracterizan un proceso biológico en una población celular.

La herramienta fue desarrollada en software de libre uso, por lo que el acceso a esta no se encuentra restringido a licencias comerciales o a incompatibilidades de versiones. De este modo, se garantiza que cualquier tipo de usuario pueda usar la herramienta, y dado el caso de poseer los conocimientos, pueda modificarla en caso de encontrar errores o mejorar los algoritmos ya implementados.

Durante la verificación de las unidades correspondientes a los modelos implementados para la inferencia de parámetros, se evidenció que esta última es una tarea compleja que requiere de conocimiento previo sobre el sistema. Esto se debe a que el proceso de inferencia requiere del uso de algoritmos de optimización, los cuales son sensibles a los valores iniciales usados en la función objetivo del sistema. Por lo tanto, es común que cuando un sistema sea complejo o cuente con muchas dimensiones, el algoritmo de optimización se quede estancado en óptimos locales o incluso sea incapaz de localizar un óptimo.

Entre las limitaciones de la herramienta se encontró que la simulación poblaciones celulares de alrededor de cientos de miles de células conlleva la manipulación de grandes cantidades de datos, lo cual representa un consumo considerable de memoria y trabajo del procesador. Esto puede llevar a que, en ciertas ocasiones, aunque el código fuente se siga ejecutando, la interfaz gráfica de la herramienta se congele. De igual manera ocasiona que el número de expresiones que puedan ser trazados en una gráfica sea limitado.

La experiencia educativa, mediante la cual se evaluó la usabilidad de la herramienta, llevada a cabo con el desarrollo de guías y un taller práctico con estudiantes de ingeniería biomédica de la Universidad Autónoma de Occidente permitió resaltar la utilidad y la pertinencia de la herramienta en cuanto al proceso de aprendizaje sobre el modelado de sistemas biológicos en poblaciones celulares. Este tipo de herramientas podrían representar un punto de partida para analizar un sistema biológico, y plantear interrogantes que puedan ser abordados de forma computacional o experimental.

10. RECOMENDACIONES

Como trabajo futuro, se plantea la implementación de algoritmos de optimización estocásticos. Esto con el fin de mejorar el rendimiento de la inferencia de parámetros en problemas no lineales de alta dimensionalidad, los cuales pueden contener múltiples óptimos locales en donde los algoritmos de optimización determinísticos pueden quedarse atrapados. Por el contrario, los algoritmos estocásticos proveen un enfoque alterno que permite tomar en cuenta menos decisiones de óptimos locales durante el procedimiento de búsqueda, aumentando de este modo, la probabilidad de localizar el óptimo global de la función objetivo (Brownlee, 2021).

Adicionalmente, sería optimo indagar e implementar otros métodos numéricos que permitan solucionar un sistema de ecuaciones diferenciales de forma más rápida. Durante el desarrollo del proyecto se probaron diferentes solucionadores de la suite de *Scipy*, los cuales, aunque proveen un desempeño adecuado para los propósitos de la herramienta, su principal limitación radica en que su integrador básico Runge-Kutta se encuentra escrito directamente en ciclos de Python, al igual que el sistema de ecuaciones diferenciales a solucionar. Por ejemplo, en (Rackauckas, 2017) realizan una comparación detallada de diferentes suites de solucionadores de ecuaciones diferenciales, entre los que destaca CVODE de *Sundials*. Esta suite provee un solucionador que de manera eficiente implementa los métodos de Adams y BDF en C y Fortran. De este modo, adicionar a la herramienta integradores de ODEs compilados en lenguajes de programación de bajo nivel; permitiría obtener tiempos de ejecución de código inferiores en comparación a los alcanzados en Python.

REFERENCIAS

Almquist, J. (2017). Kinetic Models in Life Science. https://publications.lib.chalmers.se/records/fulltext/253263/253263.pdf

Barizien, A. (2019). *Studying the variability of bacterial growth in microfluidic droplets* (Doctoral dissertation).

Brownlee, J. (2021). A Gentle Introduction to Stochastic Optimization Algorithms [en línea]. Obtenido de https://machinelearningmastery.com/stochastic-optimization-for-machine-learning/.

Cao, Y., Li, H., y Petzold, L. (2004). Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, *121*(9), 4059-4067.

Cinquemani, E. (2019). Identification, estimation and control of gene expression and metabolic network dynamics (Doctoral dissertation, Université Grenoble-Alpes, ED MSTII).

Duveau, F., Hodgins-Davis, A., Metzger, B. P., Yang, B., Tryban, S., Walker, E. A., ... y Wittkopp, P. J. (2018). Fitness effects of altering gene expression noise in Saccharomyces cerevisiae. *Elife*, *7*, e37272.

El Samad, H., Khammash, M., Petzold, L., y Gillespie, D. (2005). Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, *15*(15), 691-711.

Filippi, S., Barnes, C. P., Kirk, P. D., Kudo, T., Kunida, K., McMahon, S. S., y Stumpf, M. P. (2016). Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. *Cell reports*, *15*(11), 2524-2535.

Gillespie, D. T., (1977, May). Exact Stochastic Simulation of Coupled Chemical Reactions. The Journal of Physical Chemistry, Vol. 81, No. 25, 1977.Gillespy2 1.5.7. pypi.org. Python software foundation © 2021.

Gonzalez, A. M., Uhlendorf, J., Schaul, J., Cinquemani, E., Batt, G., & Ferrari-Trecate, G. (2013, July). Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference. In *2013 European Control Conference (ECC)* (pp. 3652-3657). IEEE.

Gonzalez, A. M. (2014). Modeling of biological processes in cell populations. Doctoral dissertation, University of Pavia.

González, A. M., Cinquemani, E., y Ferrari, G., (2016). Validation methods for population models of gene expression dynamics. *IFAC-Papers Online*, *49*(26), 114-119.

Higham, D. J., (2008). Modeling and simulating chemical reactions. Society for Industrial and Applied Mathematics. SIAM REVIEW. Vol. 50, No. 2, pp. 347 – 368.

Ilie, S., Enright, W. H., y Jackson, K. R., (2009). Numerical solution of stochastic models of biochemical kinetics. CANADIAN APPLIED, MATHEMATICS QUARTERLY, Vol 17, No 3.

Janzén, D., (2012). Standard Two-Stage and Nonlinear mixed effect modelling for determination of cell-to-cell variation of transport parameters in Saccharomyces cerevisiae.

Klipp, E., Liebermeister, W., Wierling, C., y Kowald, A. (2016). *Systems biology: a textbook*. John Wiley & Sons.

Kitano, H., (2002). Systems Biology: A Brief Overview. *Systems Biology: The Genome, Legome, and Beyond.* Vol. 295.

Lavielle, M. (2014). Mixed effects models for the population approach: models, tasks, methods, and tools. Chapman and Hall/CRC.

Lei, J., (2011). Stochastic modeling in systems biology. Zhou Pei-Yuan Center for Applied Mathematics.

Liebermeister, W. (2012). *Construction and control analysis of biochemical network models* (Doctoral dissertation, Humboldt-Universität zu Berlin).Lixoft ® 2021.

Llamosi, A., Gonzalez-Vargas, A. M., Versari, C., Cinquemani, E., Ferrari-Trecate, G., Hersen, P., y Batt, G. (2016). What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. *PLoS computational biology*, *12*(2), e1004706.

Loos, C., y Hasenauer, J. (2019). Mathematical modeling of variability in intracellular signaling. *arXiv preprint arXiv:1904.08182*.

Marguet, A., Lavielle, M., y Cinquemani, E. (2019). Inheritance and variability of kinetic gene expression parameters in microbial cells: modeling and inference from lineage tree data. *Bioinformatics*, *35*(14), i586-i595.

Maruthi, L. R., Tkachev, I., Carta, A., Cinquemani, E., Hersen, P., Batt, G., y Abate, A. (2014, November). Towards real-time control of gene expression at the single cell level: a stochastic control approach. In *International Conference on Computational Methods in Systems Biology* (pp. 155-172). Springer, Cham.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90-100.

Pájaro Diéguez, M. (2017). *Mathematical modelling, analysis and numerical simulation for a general class of gene regulatory networks* (Doctoral dissertation, Programa de doutoramento en Métodos Matemáticos e Simulación Numérica en Enxeñaría e Ciencias Aplicadas (RD 99/2011)).

Rackauckas, C. (2017). A Comparison between Differential Equation Solver Suites in MATLAB, R, Julia, Python, C, Mathematica, Maple, and Fortran [en línea]. Obtenido de https://www.stochasticlifestyle.com/comparison-differential-equation-solver-suites-matlab-r-julia-python-c-fortran/.

Ruess, J., y Lygeros, J. (2013, December). Identifying stochastic biochemical networks from single-cell population experiments: A comparison of approaches based on the Fisher information. In *52nd IEEE Conference on Decision and Control* (pp. 2703-2708). IEEE.

Schnoerr, D., Sanguinetti, G., y Grima, R., 2015. Comparison of different moment-closure approximations for stochastic chemical kinetics. The journal of chemical physics 143, 185101.

Shlens, J. (2014). Notes on Kullback-Leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*.

Singh, A., y Soltani, M. (2013). Quantifying intrinsic and extrinsic variability in stochastic gene expression models. *Plos one*, *8*(12), e84301.

Stan, G. B. (2017). Modelling in biology. *Imperial College London*.

Toni, T., y Tidor, B. (2013). Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS computational biology*, *9*(3), e1002960.

Uhlendorf, J., Bottani, S., Fages, F., Hersen, P., & Batt, G. (2011). Towards real-time control of gene expression: Controlling the hog signaling cascade. In *Biocomputing 2011* (pp. 338-349).

Wang, D., Stapor, P., y Hasenauer, J. (2019). Dirac mixture distributions for the approximation of mixed effects models. *IFAC-PapersOnLine*, *52*(26), 200-206.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., y Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, *109*(21), 8340-8345.

ANEXOS

Anexo A. Documentación técnica. (Ver archivo adjunto)

Anexo B. Taller experiencia educativa. (Ver archivo adjunto)

Anexo C. Solución de otros casos de estudio. (Ver archivo adjunto)

Anexo D. Encuesta para evaluar la usabilidad de la herramienta. (Ver archivo adjunto)

Anexo E. Quiz para evaluar la apropiación de conocimientos durante la experiencia educativa. (Ver archivo adjunto)