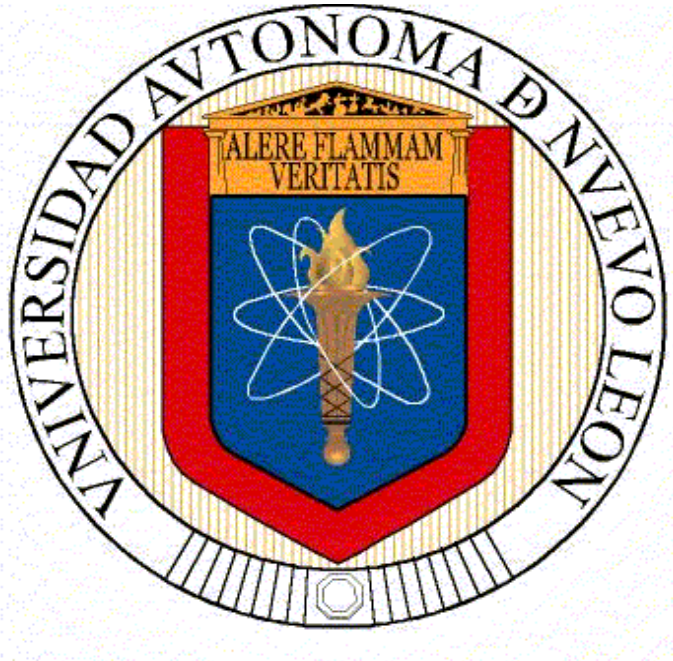
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE INGENIERGÍA MECÁNICA Y ELÉCTRICA



Inferencia de Parámetros en Líneas Celulares de Cáncer

Por

Brenda Aide Peña Cantu

En opción al grado de
Maestría en Ciencias en Ingeniería de Sistemas

Mayo 2014

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Inferencia de Parámetros en Líneas Celulares de Cáncer

por

Brenda Aide Peña Cantu

en opción al grado de

Maestría en Ciencias

en Ingeniería de Sistemas

San Nicolás de los Garza, Nuevo León Mayo 2014

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis «Inferencia de Parámetros en Líneas Celulares de Cáncer», realizada por el alumno Brenda Aide Peña Cantu, con número de matrícula 1338639, sea aceptada para su defensa como opción al grado de Maestría en Ciencias en Ingeniería de Sistemas.

El Comité de Tesis

Dr. Arturo Berrones Santos
Asesor

Dr. Romeo Sánchez Nigenda
Revisor

Dr. Víctor Treviño Alvarado
Revisor

Vo. Bo.

Dr. Simón Martínez Martínez
División de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, Mayo 2014

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

División de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis «Inferencia de Parámetros en Líneas Celulares de Cáncer», realizada por el alumno Brenda Aide Peña Cantu, con número de matrícula 1338639, sea aceptada para su defensa como opción al grado de Maestría en Ciencias en Ingeniería de Sistemas.

El Comité de Tesis

Dr. Arturo Berrones Santos
Asesor

Dr. Romeo Sánchez Nigenda
Revisor

Dr. Víctor Treviño Alvarado
Revisor

Vo. Bo.

Dr. Simón Martínez Martínez
División de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, Mayo 2014

A mi tía Elisa Martha y
a mi tío José Salvador.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Principalemte agradezco a Dios por permitirme finalizar una etapa más en esta aventura de la vida. Agradezco al gran ángel de la guarda quien me cuida y me vigila desde asientos VIP, mi mamá. Le agradezco por todos los sacrificios, consejos y enseñanzas que me dejo.

Quiero agradecer especialmente a mi tia Marianela, quien me abrio las puertas de su hogar y de su corazón. Además agradezco a mi padre por su presencia. A mi hermano Andrés, a mis hermanas Nora, Marina, Rita y Rosy. A mis queridos sobrinos Alan, Abrham, Gustavo, Marinee, Oziel y Andrea.

Agradezco todo el apoyo mostrado por mis tios Salvador, Martha, Diana, Irasema, Hector, Fernando, Martha Nelly e Yvonne. A todos mis primos que definitivamente tienen un lugar en mi vida y en mi corazón.

Doy gracias a Ashanti, mi compañera de viaje en esta vida. Mi motivación e inspiración. Gracias por estar conmigo en las buenas y en las malas. Gracias por todos esos momentos que hemos compartido, por todos los buenos recuerdos que disfruto a tu lado, por ser la persona con la que cuento en mis tristezas y una vez más en mis alegrías. Te agradezco por ser quien eres, por quedarte a mi lado y por creer en mi.

Además quiero agradecer a la familia que he elegido, mis amigos. Quiero darles las gracias a Paola, Gris, Sylvia, Lupita y a toda la banda del grupo 13. A mis grandes amigos de la banquita Beli, Diana, Clari, Carlos, Neto, Rafa y Ray. A mi amiga y colega Alma que aunque estes lejos, para mi siempre estas cerca.

Le agradezco a todos los del PISIS, tanto profesores como alumnos. Ya que me llevo recuerdos de cada uno de ellos. En especial le doy las gracias a mis amigos y compañeros Nancy, Juan, Luis, Lilian, Christopher, Dago, Liliana, Dory, Ruth y David. Gracias por compartirme su amistad, sus conocimientos, sus preocupaciones, sentimientos y su cariño. A la banda de Yalma, a Fernando, Cristina, Nancy, Paulina y Nelly por aquellas charlas tan amenas y tan graciosas.

Le doy gracias al Doctor Arturo Berrones por su guía y dirección en este proyecto. Por su gran paciencia y sus extraordinarios consejos. Además le agradezco mucho al Doctor Javier Almaguer por su ayuda y sus brillantes contribuciones al estudio que realizamos. A Edgar Jímenez le agradezco muchisimo por el apoyo, ideas, aguante y dedicación que aporto a nuestro proyecto.

Agradezco mucho a los doctores Victor Treviño y Romeo Sánchez por formar parte de mi cómite de Tesis ya que han participado de forma positiva en la revisión y enriqiecimiento de este proyecto.

También agradezco a CONACyT por brindarme, a través de su beca, la valiosa oportunidad de cursar una etapa mas que ha enriquecido satisfactiriamente en mi persona. Así mismo agradezco a mi *Alma Mater*, la Universidad Autónoma de Nuevo León, por abrirme sus puertas una vez más a esta inolvidable experiencia. Por no hacerme sentir como una alumna mas, sino como miembro de una gran familia. Además le doy gracias a mi querida Facultad de Ingeniería Mecánica y Eléctrica por su contribución en mi formación tanto profesional como personal; por fomentar y alentar la participación en Congresos y Seminarios dentro y fuera de nuestra institución. Y muy especialmente le doy gracias por su apoyo brindado a través de su beca interna y de rectoría.

Resumen

Brenda Aide Peña Cantu.

Candidato para el grado de Maestro en Ingeniería
con especialidad en Ingeniería de Sistemas.

Universidad Autónoma de Nuevo León.
Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio:

Inferencia de Parámetros en Líneas Celulares de Cáncer

Número de páginas: 46.

Objetivos y método de estudio: En el presente trabajo se muestra la comparación de métodos formulados para dar solución a la inferencia de parámetros en redes de regulación genética artificiales. Este proyecto es motivado por un caso práctico en una red genética proveniente de líneas celulares de cáncer reales al que desea extenderse.

Dentro de este estudio se incluyen evaluaciones empíricas donde se compara el desempeño de cada una de las metodologías presentadas tomando diversas muestras de datos mediante los modelos creados.

El propósito de esta tesis es contrastar distintas metodologías desarrolladas

para la inferencia de parámetros de redes genéticas artificiales. Esto con el fin de proporcionar evidencia sobre cuál es más apropiada emplear, basandose en criterios de eficiencia y errores de muestra.

Contribuciones y conlusiones: La contribución fundamental del presente trabajo radica en realizar un análisis de las metodologías creadas para inferir parámetros teniendo en cuenta la limitación de pocos datos atribuida a las escasas observaciones con las que se cuentan en experimentos en casos reales.

Firma del asesor: _____

Dr. Arturo Berrones Santos

Capítulo 1

Introducción

El problema abordado en esta tesis proviene de la necesidad de comprender el diseño y funcionamiento de organismos vivos desde una perspectiva microscópica. Lo cúal se puede lograr con la ayuda de la biología de sistemas, que consite en el estudio de un organismo o sistema biológico.

La biología de sistemas es una rama joven de la ciencia que integra diversas disciplinas como la biología, química, física, medicina, ingieniería, matemáticas, entre otras. Esto con el objetivo de obtener un mayor entendimiento de los sistemas vivos y de sus procesos biológicos. Esta disciplina representa una estrategia análitica que permite relacionar los elementos de un sistema, con el objetivo de comprender sus propiedades emergentes. En general, un sistema puede estar compuesto por tan solo unas cuantas moléculas de proteínas o unos cuantos genes que realizan una serie de actividades. Conjuntamente forman parte de una maquinaria molecular más compleja, o un grupo de células que ejecutan una función concreta.

Por lo descrito anteriormente, el análisis de sistemas puede aplicarse a moléculas, células, órganos, individuos o incluso ecosistemas completos. Siendo clave el manejo de poca información disponible como es el caso en el estudio y análisis genético realizado en el presente trabajo. Obteniendo como resultado valiosas herramientas estadísticas para deducir valores y parámetros en procesos genéticos.

1.1 Descripción del Problema

Cuando se desea estudiar problemas relacionados con la medicina, es interesante conocer los retos y obstáculos que se presentan día a día en un campo tan amplio como la salud. Ya que podemos ver como poco a poco se han ido integrando conocimientos de disciplinas tan diversas como matemáticas, física, biología, informática y ciencias computacionales. Ésto con el fin de crear una sinergía en cuanto a la resolución de problemas. Por lo que en la actualidad es mucho más común atacar problemáticas desde múltiples puntos de vista.

Teniendo en cuenta lo mencionado anteriormente podemos describir este trabajo como un desafío sumamente arduo cuando hablamos de problemas biológicos. En el presente trabajo tenemos la dificultad de actuar con poca y escasa información. Además nos vemos a la tarea de desarrollar e implementar sistemas biológicos con el fin de que sirvan de base para comprobar las metodologías elaboradas.

Primeramente se tiene una motivación de un caso real en un estudio de líneas de cáncer de mama. El cual se explora la hipótesis del vínculo o la influencia que pueden ejercer los ritmos circadianos, es decir las oscilaciones de las variables biológicas en intervalos regulares de tiempo [1] en la formación de neoplasias, en particular, de tejido mamario.

Se puede decir que el trabajo mostrado en el presente escrito corresponde a un primer acercamiento de una investigación mucho más amplia. La cuál tiene como propósito analizar y entender las secuecias genéticas con el fin de encontrar características y patrones. Que a su vez éstos puedan ayudar a desenvolver el misterio de las conexiones entre genes y síntesis de proteínas, descubiendo asi, su función en procesos biológicos de interés. Por lo que el estudio realizado en este texto se puede entender como un preprocesamiento de un análisis mucho más arduo y extenso.

Para emprender los primeros pasos este problema se aterriza inicialmente en una red genética simple, que en pocas palabras podemos describirla como conex-

iones e interacciones entre genes, ésto con el fin de implementarla y trabajar en ella metodologías estadísticas de muestreo, suavización e inferencia de datos.

1.2 Motivación

El presente trabajo es motivado por un problema real presente en un grupo de investigación en Medicina del Tecnológico de Monterrey Campus Monterrey. Su interés y relevancia radica en la nececidad actual de descubrir la relación de causa efecto en enfermedades de impacto global como lo es el cáncer de mama. Esto se debe a que son procesos que ocurren dentro de cada organismo vivo y son de gran interés de estudio.

1.3 Justificación

En la actualidad se ha mostrado un crecimiento en el interés de analizar y estudiar sistemas biológicos en organismos vivos ya que tiene múltiples aplicaciones direcamente en disciplinas como medicina y biología. Por lo que es de suma disposición realizar aportaciones de gran magnitud como lo son las herramientas para inferir datos con poca información disponible, ya que el desarrollo de experimentos médicos y biológicos demanda un gran esfuerzo tanto científico, económico y social.

1.4 Hipótesis

Los métodos estadísticos para inferir parámetros desarrollados en el presente estudio establecerán valores aceptables en el análisis de datos en las redes genéticas artificiales estudiadas.

1.5 Objetivo

El propósito general de este estudio es desarrollar y validar una metodología computacional estadística que proporcione valores aceptables en una gama de redes de regulación genética artificiales. La cuál sea facilmente extendible a casos reales.

1.6 Estructura de la Tesis

En el Capítulo 2 se abordan los antecedentes y marco teórico que sirven para el desarrollo de este trabajo, además se presentan detenidamente conceptos básicos para comprender el problema de estudio. En el Capítulo 3 se exponen las características y propiedades del modelo estudiado, se explica detalladamente la metodología de estudio, asimismo las aplicaciones y extensiones de ésta. En el Capítulo 4 se presenta el trabajo empírico sobre la comparación de las metodologías mostradas y se exponen los resultados obtenidos en base a la experimentación. En el Capítulo 5 se establecen las conclusiones y se comenta el trabajo futuro. Por último el Capítulo **??** corresponde al Apéndice del texto y en él se complementan los resultados gráficos del Capítulo 4.

Capítulo 2

Marco Teórico

La inferencia de datos es un proceso mediante el cual se realizan estimaciones sobre parámetros de relevancia a partir de un conjunto de observaciones que provienen de un sistema. En el presente trabajo se aborda una metodología en especial, conocida como inferencia Bayesiana. Los métodos de la teoría de inferencia Bayesiana fueron desarrollados en el siglo XIX. Donde esta teoría propone funciones complejas que en la mayoría de la veces no pueden ser resueltas analíticamente.

En la actualidad se ha retomado poco a poco el interés en la teoría Bayesiana ya que tiene la capacidad de aplicarse en muchos procesos y sistemas reales, esto debido a los avances y desarrollos tecnológicos. Por tales motivos es que los métodos de inferencia Bayesiana poseen un gran potencial para atacar procesos biológicos como lo son redes de regulación genética.

En el presente capítulo se muestran las bases y enfoques de las metodologías desarrolladas con el fin de tener un panorama amplio de conocimientos y herrameintas para atacar poroblemas inspirados en la medicina y biología.

2.1 Biología de Sistemas

En la actualidad existen nuevas técnicas que dan lugar al acceso de miles de datos y mayor poder computacional. Algoritmos nuevos han cambiado la visión de muchos científicos sobre cómo solucionar problemas desconocidos y retadores. Ésto

ha dado lugar a una nueva disciplina conocida como Biología de Sistemas, la cual fusiona en una estructura más elemental a expertos en áreas tan diversas como Biología, Matemáticas, Física, Informática y Medicina, entre otras [16].

La Biología de Sistemas se fundamenta en el estudio de un organismo o sistema biológico, visto como un sistema integrado e interrelacionado de genes, proteínas y reacciones bioquímicas que dan lugar a procesos biológicos. Lo que intenta hacer esta disciplina es estudiar los componentes de un organismo y sus interacciones como un solo sistema en lugar de analizar los componentes individualmente. Por lo tanto esta disciplina representa una estrategia análitica para relacionar elementos de un sistema, con el objetivo de integrar sus propiedades emergentes. Por ejemplo, un sistema biológico puede estar compuesto por solo unas móleculas de proteínas que realizan una síntesis de ácidos grasos, formando parte de una maquinaria molecular como los es una transcripción, o un grupo de células que ejecutan una función concreta, como la respuesta inmune [16]. Por lo tanto, el análisis de sistemas puede aplicarse a moléculas, células, órganos, individuos o incluso ecosistemas.

Es importante saber que el objetivo de la Biología de Sistemas radica en integrar información confiable con el fin de que se consiga un mayor entendimiento de las interacciones entre los compomentes de los sistemas vivos y por consiguiente de sus procesos biológicos. Con el fin de alcanzar dicho objetivo se desarrollan herramientas como modelos matemáticos, simulaciones y técnicas de procesamiento de datos que complementan las estrategias empíricas actuales de las ciencias biológicos.

Las principales tecnológias empleadas en la Biología de Sistemas son tanto técnicas experimentales como técnicas computacionales. Dentro de las técnicas experimentales podemos nombrar las siguientes:

- Análisis de la secuencia genética.
- Análisis de de expresión genética.
- Análisis de interacciones ADN-proteínas.

- Análisis de interacciones proteína-proteína.
- Análisis de localización subcelular proteica.

Entre las técnicas computacionales empleadas por la Biología de sistemas tenemos:

- Desarrollo de algoritmos.
- Agrupaciones de datos.
- Simulación de proceso biológicos.
- Desarrollo de modelos matemáticos.
- Diseño y análisis de sistemas.

Es fundamental comprender el comportamiento del sistema y en que medida afectan los cambio externos a este y de que forma hay que responder y modificarlo para adaptarse a dichos cambios.

El análisis teórico de sistemas biológicos mediante modelos matemáticos se ha utilizado ampliamente en el estudio de ecosistemas y de dinámica de poblaciones. De la misma manera, en la Biología de Sistemas el objetivo del análisis téorico es utilizar lenguaje matemático para describir el comportamiento de los sistemas biológicos. Las principales herramientas usadas en la Biología de Sistemas son aquellas que en Física se utilizan para estudiar sistemas dinámicos complejos, como ecuaciones diferenciales, análisis de bifurcaciones, análisis de balance de flujo, por mencionar algunos [16]. Entre los modelos matemáticos comúnmente utilizados en Biología de Sistemas están [4]:

- Modelos estadísticos.
- Modelos cinéticos.

- Redes neuronales o modelos de Markov.
- Modelos metabólicos.

Otra herrameinta importante en la Biología de Sistemas son las simulaciones computacionales, ya que juegan un papel sumamente importante porque su principal objetivo es reproducir de forma correcta el comportamiento de un sistema biológico. En dichas simulaciones se integran datos experimentales obtenidos mediante análisis teórico.

Las diferentes herramientas computacionales propias de la Biología de Sistemas permiten el estudio de las interacciones biológicas que se producen a un determinado nivel, es decir de las redes biomoleculares dentro de las que encontramos:

- Redes metabólicas.
- Redes transcripcionales.
- Redes de regulación genética.

Esta área nos permite comprender los mecanismos biológicos ya que proporciona una forma de representar los componentes celulares y sus interrelaciones y permite la identificación y caracterización de módulos funcionales. El estudio de las redes celulares se puede realizar mediante dos estrategias bien diferenciadas, auque poseen la misma finalidad, se acercan al problema desde puntos de vista opuestos.

La primera estrategia consiste en abordar directamente las redes para descomponerlas poco a poco en subredes de menor complejidad, que forman agrupamiento de proteínas. De esa manera se puede obtener información como nuevas rutas metabólicas, asi como interacciónes desconocidas entre proteínas.

La segunda estrategia se enfoca en estudiar interacciones conocidas, pero al mismo tiempo caracterizarlas poco a poco en estructuras complejas macromoleculares mediante técnicas de alta resolución. El propósito de esta estrategia consiste

en reconstruir virtualmente redes celulares e incluso una célula completa a partir de complejos proteicos [2, 16].

2.1.1 Redes de Regulación Genética

Una red de regulación genética consiste en una colección de segmentos de ADN en una célula que interactúan entre si directa o indirectamente. Esta interacción se da entre su RNA y su expresión de proteínas. Habitualmente la molécula de RNAmensajero va a constituir una proteína o conjunto de proteínas específicas [14].

La represión y la inducción son mecanismos mediante los cuales se modula la expresión de los genes y las moléculas que intervienen en los procesos biológicos. Se puede ver a los genes constituir la red tal como si fueran nodos, cuyas entradas son las proteínas como factores de transcripción y como salida se tiene el nivel de expresión genética. Un gen o nodo puede ser visto como una función que se obtiene de la combinación de funciones sobre las entradas. Asimismo las concentraciones de proteínas actuan como controladores fundamentales dentro de las células ya que determinan las coordenadas espaciales y temporales de la célula [7].

Es interesante saber que las redes de regulación genética se encuentran en la etapa de entender el comportamiento del sistema en niveles crecientes de complejidad. Los niveles de expresión de un gen están determinados por la transcripción de su RNAmensajero y por la traducción de éste en proteínas. La existencia de concentraciones diferentes de proteínas codificadas explicaría la regulación a nivel de transcripción de ADN a RNAmensajero.

2.2 EL Represilador

El modelo principal de estudio del presente trabajo es conocido como *Represilador* introducido y estudiado por primera vez por Elowitz y Leiber en el 2000 [10]. El cual consiste en en una red sintética de 3 genes transcripcionales de E. coli

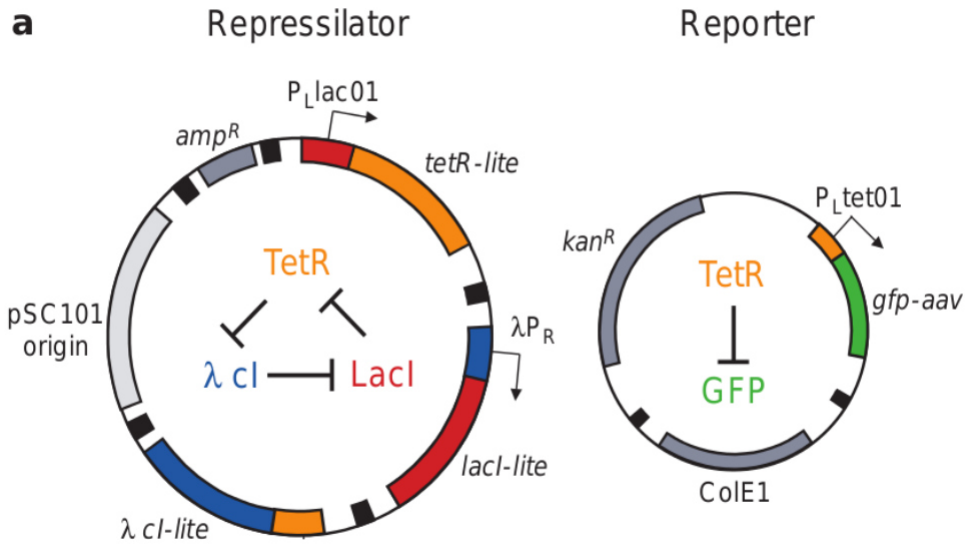


Figura 2.1: Red sintética transcripcional de E. coli

[22] donde se induce periodicamente la síntesis de proteína verde fluoresente (GFP, por sus siglas en inglés) como lectura del estado de las células individuales. En las Figuras 2.1 y 2.2 podemos observar la red descrita anteriormente.

Esta red de regulación genética ha sido múltiples veces estudiada y analizada en los últimos años [5, 15, 18] ya que posee características interesantes de las cúales se pueden aprender y apreciar un sinfín de procesos diversos.

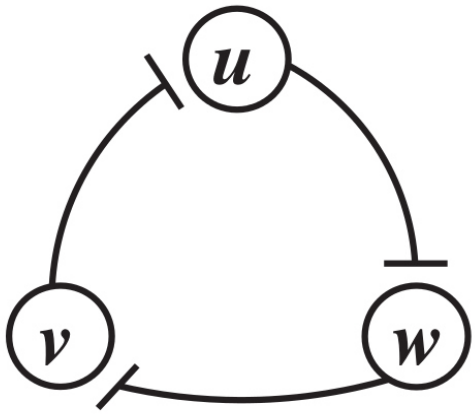


Figura 2.2: Red sintética transcripcional de E. coli

2.3 Métodos de muestreos

Cuando se realiza una introducción general de la estadística se dice que uno de los objetivos fundamentales es obtener conclusiones basadas en los datos que se han observado. A dicho proceso se le conoce como inferencia estadística, es decir mediante la información recabada por una muestra de la población se obtienen conclusiones o se infieren valores sobre características poblacionales [19].

En el análisis de procesos biológicos es común contar con pocas muestras, o pocos datos observados. Es por ello que frecuentemente se hace uso de técnicas de muestreo en el esfuerzo de estudiar procesos bilógicos. La estadística reconoce al muestreo como la técnica para la selección de una muestra a partir de una población. Este proceso ahorra recursos y a su vez obtiene resultados muy parecidos a los que se llegaría si se reaizara un estudio con toda la población [19].

En general existen dos métodos para seleccionar muestras de poblaciones:

- Aleatorio.

- No aleatorio.

El muestreo aleatorio cumple con la condición de que todos los elementos de la población tienen probabilidad de ser escogidos. En cambio el muestreo no aleatorio se basa en la experiencia de alguien sobre la población. Existen múltiples técnicas que se pueden usar para llevar a cabo un muestreo aleatorio como:

- Muestreo simple.
- Muestreo sistemático.
- Muestreo estratificado.
- Muestreo por estados múltiples.
- Muestreo por conglomerados.

En trabajo realizado en el presente texto se hizo uso de los primeros dos tipos de muestreos, muestreo aleatorio simple y muestreo aleatorio sistemático. Un muestreo aleatorio simple consiste en seleccionar un tamaño de muestra n de una población N, de tal manera que cada muestra posible de tamaño n tenga la misma probabilidad de ser seleccionada [6]. Por otro lado un muestreo aleatorio sistemático tiene lugar cuando se tiene un tamaño de muestra n de una población N, donde además se tiene una constante de elevación k que consiste en $k = N/n$, la cual determina el intervalo en el que se realiza la extracción [6].

2.4 Suavizado de datos

En la estadística se conoce el suavizado de datos como crear una función que tiene como finalidad capturar los patrones importantes del conjunto de datos, dejando de lado el ruido [21]. Para ello se hace uso de diversos algoritmos, siendo los más comúnes el de promedios móviles, exponencial, exponencial doble y el de tendencia y estacionalidad.

2.5 Simulación

Uno de los puntos fundamentales de la presente tarea es el de la simulación de proceso biológicos. Ya que es de sumo valor e interés científico. Una simulación consiste en investigar una hipótesis o conjunto de hipótesis de trabajo ayudandose de modelos. Asimismo se define como una técnica numérica para conducir experimentos en una computadora, donde estos experimentos conducen cierto tipo de relaciones matemáticas y lógicas, las cuales describen el comportamiento y estructura de sistemas complejos del mundo real [11]. Además también puede definirse como el proceso de diseñar un modelo de un sistema real y llevar a cabo experiencias con el, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias para el funcionamiento del sistema [20].

2.6 Teoría Bayesiana

En algún momento de la década de 1720, el reverendo Thomas Bayes realizó un ingenioso descubrimiento comenzando a estudiar diversas formas de abordar matemáticamente la cuestión de causa y efecto. A principios del siglo *XVIII* apenas existían nociones de análisis probabilístico por lo que el único ambito en el que los cálculos se aplicaban extensamente era el de los juegos de azar, el cual era utilizado para abordar cuestiones básicas como las probabilidades de conseguir cuatro ases en una mano de póquer. Sin embargo, nadie había logrado averiguar la forma de dar la vuelta a las deducciones y retroceder en la secuencia causal a fin de poder plantear la pregunta inversa, la que asendía del efecto a su causa. ¿Qué ocurre si un jugador de póquer se reparte a sí mismo cuatro ases en tres manos consecutivas?, ¿Cúal es la probabilidad subyacente (o la causa) de que se consiga una triple partida de semejantes características? [3].

No se sabe con exactitud qué fue lo que despertó el interés de Bayes en el problema de la probabilidad inversa, pero una vez cristalizada en su mente la esencia de

ésta, decidió que su objetivo pasaba por determinar la probabilidad aproximada de un acontecimiento futuro del que no tuviese información alguna salvo la derivada de sus circunstancias pasadas, esto es, la vinculada con el número de veces que dicho acontecimiento hubiera tenido lugar o, por otro lado, hubiera dejado de producirse. Para cuantificar el problema, Bayes necesitaba una cifra, y en algún momento entre los años 1746 y 1749 finalmente logró dar con una solución ingeniosa. Reduciendo el problema a sus elementos más básicos, él imaginó una mesa cuadrada tan plana y bien nivelado que al hacer rodar una pelota sobre ella ésta tuviera tantas probabilidades de ir a parar a en punto A como de terminar en otro punto B.

Con lo descrito anteriormente, él sentado de espaldas a la mesa para no ver nada comienza pidiendo que se eche a rodar una imaginaria bola blanca sobre la superficie de la mesa. Posteriormente pide que se haga rodar una segunda bola sobre la mesa y que le informen si ésta viene a parar a la izquierda o a la derecha de la bola blanca. En caso de que lo haga a la izquierda, se comprenderá que hay más probabilidades de que la bola blanca se encuentre en la parte derecha de la mesa. De nuevo se impulsa la bola y se reporta únicamente si ésta se detiene a la derecha o a la izquierda de la bola de prueba. Si lo hace a la derecha, él inferirá que la bola blanca no puede hallarse al borde derecho de la mesa. Bayes va pidiendo que se ponga a rodar una y otra vez, la bola del experimento. Los jugadores y matemáticos ya sabían que cuantas más veces lanzaran una moneda al aire, más fiables serían sus conclusiones. Lo que Bayes descubrió fue que al aumentar el número de bolas que se echaban a rodar por la mesa, cada nuevo dato registrado hacía que las oscilaciones del punto de asiento de su imaginaria bola blanca de referencia se movieran en un área cada vez más restringida.

Por ejemplo, un caso extremo sería si todas las bolas lanzadas después de la primera se detubieran a la derecha de ésta, en dicho caso se tendría que concluir que los más probable era que la bola blanca estubiese situada en el extremo marginal izquierdo de la mesa. Por otro lado, si todos los lanzamientos quedaran a la izquierda de la primera bola, lo más probable sería que esta se encontrara en el borde derecho.

Al final, suponiendo que se hubiese lanzado la bola un número de veces suficiente, Bayes podía reducir progresivamente el área de la posible ubicación de la primera bola lanzada.

La genialidad del experimento de Bayes radicaba en el hecho de concebir la idea de estrechar la gama de posibles posiciones de la bola inicial en inferir basándose en tan escasa información que se hallaba detenida en algún punto situado entre dos límites concretos. Este enfoque era incapaz de generar una respuesta correcta. Bayes nunca tendría la posibilidad de saber con precisión el desplazamiento exacto de la bola blanca, pero podría afirmar con progresiva confianza que lo más probable era que se encontrara inscrita en un determinado espacio. De tal modo, el sencillo y limitado sistema diseñado por Bayes pasaba de las observaciones del mundo a su origen o causas probables. Valiéndose de este conocimiento del presente (es decir, de la información sobre las posiciones de las bolas lanzadas, bien a la derecha, bien a la izquierda de la bola de control). Hasta le resultaba posible valorar el grado de confianza que podía depositar en la conclusión a la que hubiese llegado.

Desde el punto de vista conceptual, el sistema de Bayes era extremadamente simple, es decir, modificamos nuestras opiniones al recibir una información objetiva:

Creencias de partida (primera conjetura vinculada con la posible posición de la bola de control) + Los datos objetivos recientes (si la ultima bola ha ido a la izquierda o a la derecha de la conjetura inicial) = Creencia nueva y mejrada.

Al final se le darían nombres a las distintas partes de este método:

- *A priori* : Probabilidad de la creencia inicial.
- *Verosimilitud* : Grado de probabilidad de las sucesivas hipótesis construidas sobre la base de los nuevos datos objetivos.
- *A posteriori*: Probabilidad de la creencia recien revisada.

Cada vez que se efectúa un nuevo cálculo, la probabilidad *a posteriori* se con-

vierte en la *a priori* de la nueva repetición. Por lo tanto se trata de un sistema que va evolucionando, de tal modo que cada nuevo aporte de información va aproximando cada vez más a la certidumbre del experimentador. En resumen:

El *a priori* multiplicado por la *verosimilitud* es proporcional al *a posteriori* [3].

Habiendo introducido el origen detrás de la teoría de Bayes, es importante describirla formalmente. Por lo que si tenemos una muestra aleatoria $X_1, ..., X_n$ de una población siendo su función de probabilidad de masa (pmf, por sus siglas en inglés) o su función de densidad de probabilidad (pdf, por sus siglas en inglés) [9] $f(x; \vartheta)$, donde $x \in \chi$ y $\vartheta \in \Theta$ y el parámetro no conocido v se supone fijo. Una inferencia frecuentista, es decir, una simple inferencia de hipótesis, produce dependencia en la función de *verosimilitud*, denotada como $L(\vartheta) = \Pi_{i=1}^{n} f(x_i; \vartheta)$, donde ϑ es desconocido y fijo.

Hablando del enfoque bayesiano, el experimentador debe tener en mente desde el principio que el parámetro ϑ es una variable aleatoria teniendo su propia distribución de probabilidad en el espacio Θ. Ahora que ϑ es aleatoria, la función de *verosimilitud* será la misma que $L(\theta)$ dado que $\vartheta = \theta$. Por lo que se denota la pmf o pdf de ϑ por $h(\theta)$ en el punto $\vartheta = \theta$, la cual se conoce como distribución *a priori* de ϑ.

La distribución *a priori* refleja una "creencia" subjetiva con respecto a cuáles valores de de v son los más probables considerando todo el espacio de parámetros Θ. La distribución *a priori* debe estar bien fijada antes de que se realize la recolección de datos. En este caso el experimentador puede guiarse por su experiencia o conocimientos de tal manera que se obtenga una distribución *a priori* que se apegue a la realidad.

El paradigma bayesiano debe realizar todas las inferencias y análisis después de combinar la información de ϑ contenida en toda la evidencia recolectada por la función de *verosimilitud* $L(\theta)$ dado que $\vartheta = \theta$, asi mismo que de la distribución *a priori* $h(\theta)$. Combinando la evidencia de v derivada de la distribución *a priori* como

la información de la función de *verosimilitud* podemos obtener la distribución *a posteriori*. Por lo tanto todas las inferencias Bayesianas son guiadas por su distribución posterior.

2.6.1 Distribuciones *a priori* y *a posteriori*

Como se mencionó antriormente el parámetro ϑ se asume como desconocido y como si fuera una variable aleatoria teniendo su pmf o pdf $h(\theta)$ en el espacio Θ. En este caso $h(\theta)$ es la distribución *a priori* de v. La información de v derivada de la pdf anterior es combinada con el resultado de la función de *verosimilitud*. Es importante mencionar que la función de *verosimilitud* es el conjunto condicional de la pmf o pdf de $\mathbf{X} = (X_1, ..., X_n)$ dado $\vartheta = \theta$. Si se supone que existe un T suficiente para θ en la función de *verosimilitud* observada para $\mathbf{X}$ dado que $\vartheta = \theta$. El estadístico T suficiente será frecuentemente evaluado con su pmf o pdf $g(t;\theta)$ dado que $v = \theta$, dado que $t \in T$ donde T es un subconjunto de los números reales ($\mathfrak{R}$). Si T es una variable continua y por esto las probabilidades y expectativas asociadas se describen como integrales en el espacio de T. Las integrales se interpretan como sumas en el caso en que T sea una variable discreta.

El conjunto de la pdf de T y v es dado por:

$$g(t;\theta)h(\theta) \quad \forall \quad t \in \mathcal{T} \quad y \quad \theta \in \Theta. \tag{2.1}$$

La pdf marginal de T se obtiene integrando el conjunto de la pdf 2.1 respecto a θ, es decir, podemos escribir la pdf marginal de T de la manera 2.2:

$$m(t) = \int_{\theta \in \Theta} g(t;\theta)h(\theta)d\theta \quad \forall \quad t \in \mathcal{T}. \tag{2.2}$$

De esta manera se puede obtener la pdf condicional de v dado $\mathcal{T} = t$, como se muestra en 2.3:

$$k(\theta;t) \equiv k(\theta|T=t) = g(t;\theta)h(\theta)/m(t) \; \forall \; t \in \mathcal{T} \; y \; \theta \in \Theta \; talque \; m(t) > 0. \qquad (2.3)$$

La pdf condicional $k(\theta;t)$ de ϑ dado que $T = t$ es llamada distribución *a posteriori* de ϑ.

Bajo el paradigma bayesiano, después de obtener la función de *verosimilitud* y la distribución *a priori* la distribución *a posteriori* $k(\theta;t)$ de v enfatiza como se combina la información de ϑ obtenida de dos diferentes formas, el conocimiento *a priori* y la recolección de datos. La habilidad de daptación de la función análitica final de $k(\theta;t)$ tiende a tener una dependencia muy fuerte con qué tan sencillo o difícil sea evaluar $m(t)$. En algunos casos, la distribución marginal de T y su distribución *a posteriori* pueden solo ser evaluadas numéricamente.

Capítulo 3

Metodología de Solución

El obietivo principal de este trabajo radia en aplicar métodos, específicamente métodos Bayesianos en la inferencia de datos en redes complejas, como lo son las redes de regulación genética artificiales. De igual forma se desea analizar el desempeño de los métodos propuestos en comparación con métodos de inferencia clásicos y muy utilizados como son los de recurrencia. En el presente capítulo se presentan ambos métodos de solución con fines comparativos. En la sección 3.1 se describe el método de simulación del *Represilador* empleado en este trabajo. En la sección 3.2 se introducen los métodos de muestreo de datos empleados en las redes genéticas. En la sección 3.3 se presentan los métodos de suavización usados con la infirmación muestreada. Por último en la sección 3.4 se muestran los métodos de inferencia empleados en este estudio.

3.1 Métodos de Simulación

En esta sección se describe la metodología empleada para la simulación del *Represilador*. Primeramente se procedio a simular la red, la cual recordando de la sección anterior, consiste en una red genética artificial de 3 genes transcripcionales de E. coli que producen proteína verde fluoresente como estado de salida del sistema. Mientras la concentración de uno de los genes se ve en aumento, se puede percibir que la de su gen compañero inmediato disminuye y como la concentración de éste repercute en forma positiva a su compañero siguiente. De esta manera se puede

distinguir como se encuentran conectados los 3 genes y como se forma el ciclo de represión entre ellos. Para llevar a cabo dicha tarea se recurrió a aproximar soluciones numéricas de ecuaciones diferenciales ordinarias mediante el método de Euler [29]. El cúal tiene la estructura de 3.1.

$$y_{n+1} = y_n + hf(x_n, y_n) \tag{3.1}$$

Donde f es la función obtenida de la ecuación diferencial $y' = f(x, y)$. El uso recursivo de 3.1 para $n = 0, 1, 2, ...$ produce las ordenadas $y_1, y_2, y_3, ...$ de puntos en rectas tangentes sucesivas con respecto a la curva solución en $x_1, x_2, x_3, ...$ o $x_n = x_0 + nh$, donde h es una constante y representa el tamaño de paso entre x_n y x_{n+1}. Los valores $y_1, y_2, y_3, ...$ aproximan los valores de una solución $y(x)$ del problema de valores iniciales en $x_1, x_2, x_3, ...$ Pero sin importar la ventaja que la ecuación 3.1 tenga en su simplicidad, se pierde en la severidad de sus aproximaciones.

Supongamos que tenemos la ecuación $y' = 0.1\sqrt{y} + 0.4x^2$. Para proceder a simularla mediante el método de Euler debemos linealizarla de la siguiente manera: $L(x) = y_0 + f(x_0, y_0)(x - x_0)$. Hacemos h un incremento positivo en x de tal forma que se tenga $L(x) = y_0 + f(x_0, y_0)(x + h - x_0)$ o $y_1 = y_0 + hf(x_0, y_0)$. Recordando que el método Euler es $y_{n+1} = y_n + hf(x_n, y_n)$ donde $x_n = x_0 + nh$ para $n = 0, 1, 2, 3, ...$ y realizando 4 iteraciones con un valor h de 0.1 obtenemos la tabla 3.1. Siguendo el mismo procedimiento hasta 15 iteraciones se obtendrían los valores graficados en 3.1.

x_n	y_n
1	0.80
1.10	0.62
1.20	0.47
1.30	0.35
1.40	0.25

Tabla 3.1: Iteraciones del ejemplo del método Euler

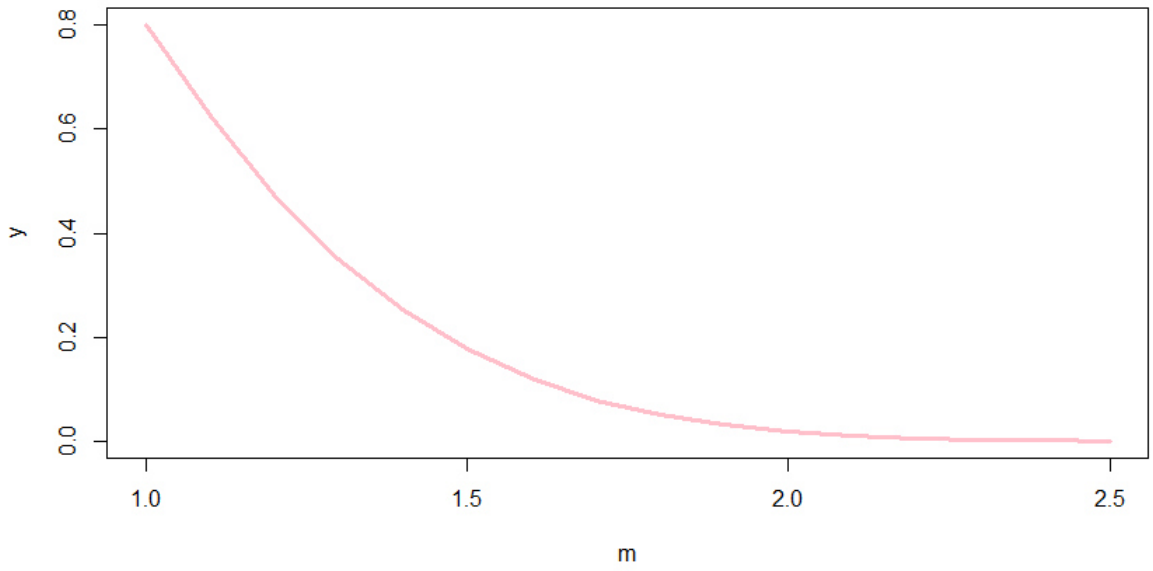


Figura 3.1: Simulación usando el método de Euler

Por otro lado, teniendo en cuenta la tarea de simular el *Represilador* tenemos la Tabla 3.2 donde se muestran las 6 ecuaciones diferenciales acopladas que forman la red descrita de la Figura 2.1, La cual se dio la labor de simular en el software libre R [24] por el método de Euler descrito anteriormente. En esta tabla se puede observar que las $m's$ corresponden a las concentraciones del RNAmensajero; u, v y

$$\begin{aligned}
\frac{dm_1}{dt} &= -m_1 + \frac{\alpha}{1+v^n} + \alpha_0 \\
\frac{dm_2}{dt} &= -m_2 + \frac{\alpha}{1+w^n} + \alpha_0 \\
\frac{dm_3}{dt} &= -m_3 + \frac{\alpha}{1+u^n} + \alpha_0 \\
\frac{du}{dt} &= -\beta(u - m_1) \\
\frac{dv}{dt} &= -\beta(v - m_2) \\
\frac{dw}{dt} &= -\beta(w - m_3)
\end{aligned}$$

Tabla 3.2: Sistema de Ecuaciones Diferenciales Acopladas del *Represilador*

w corresponden a las concentraciones de porteínas de los genes λcl, $TetR$ y $Lacl$ respectivamente. Además α, α_0, β y n representan la degradación molecular, la saturación del RNAmensajero, la razón de decaimiento de concentración de proteínas y el coeficiente de Hill respectivamente.

Posteriormente se penso y procedio a generalizar la red de Elowitz y Leibler [10] conocida como el *Represilador Generalizado* [13, 12, 23, 28]. Donde se tienen las mismas condiciones que el caso anterior solo que éste es extiendible a n número de genes. En la Tabla 3.3 podemos ver las ecuaciones diferenciales para dicha red. Donde se describen m_j y p_j como las concentraciones de RNAmensajero y proteína respectivas a cada gen j.

$$\begin{aligned}
\frac{dm_j}{dt} &= -m_j + \frac{\alpha}{1+p_{j-1}^n} + \alpha_0 \\
\frac{dp_j}{dt} &= -\beta(p_j - m_j)
\end{aligned}$$

Tabla 3.3: Sistema de Ecuaciones Diferenciales Acopladas del *Represilador Generalizado*

Además de α, α_0, β y n que al igual que en el *Reporesilador* representan la degradación molecular, la saturación del RNAmensajero, la razón de decaimiento de concentración de proteínas y el coeficiente de Hill respectivamente. Es importante mencionar que el *Represilador Generalizado* cuenta con algunas características interesantes que se discutirán el siguiente capítulo.

3.2 Métodos de muestreo

Es común recurrir a los muestreos cuando deseamos recabar información para realizar análisis estadisticos de datos. En este trabajo utilizamos dos tipos de muestreos de datos, un muestreo aleatorio simple y un muestreo aleatorio sistemático. Un muestreo aleatorio simple consiste en un muestreo equiporbabilísitco, donde se selecciona una muestra de tamaño n de una población de N unidades y cada elemento tiene una probabilidad de inclusión de n/N [8, 17, 25]. En cambio en muestreo aleatorio sistemático elegimos a los n elementos en base a un intervalo de salto definido como k, en el cual elegimos de manera aleatoria el primer elemento y seleccionamos el siguiente cada salto k [8, 17, 25].

En el presente estudio se recurren a diversos tamaños de muestra de datos. Además es importante señalar que se realizaron tanto muestreos temprorales como repeticiones. Esto con el propósito de apegarse a estudios de procesos biológicos reales, es decir se pueden obtener muestras de experimentos que se pueden repetir varias veces con el fin de recabar mayor información de los sistemas estudiados.

3.3 Métodos de suavización

Una herramienta útil en la estadística es la suavización de datos, la cual ayuda a eliminar fluctuaciones aleatorias de una serie de tiempo, dando datos menos distorsionados del comportamiento real de la misma. La idea central es definir una nueva serie a partir de la serie de tiempo creada con los datos observados, de tal

manera que se suavicen los efectos ajenos a la tendencia [26]. Es importante saber que existe una inmensa variedad de suavizaciones, siendo las más comunes la exponencial simple, exponencial doble y la suavización con tendecia y estacionalidad. En el presente estudio utilizamos cuatro tipos de suavizaciones:

- Aproximación por regresión de polinomios locales.
- Estimación de suavizaciones cúbicas.
- Función de estimación mediante polinomios locales.
- Suavización Bayesiana de una serie de Fourier.

Siendo las primeras tres funciones pertenecientes al software libre R [24] y la última implementada bajo una serie de Fourier cosenoidal. Una aproximación por regresión de polinomios locales (más adelante definida como *spline*) acopla una superficie polinómica determinada por uno o varios predictores numéricos acoplandolos de manera local [24, 26]. Por otro lado la estimación de suavizaciones cúbicas (descrita posteriormente como *loess*) se adapta a una *spline* cúbica con los datos alimentados [24, 26]. Mientras que una función de estimación mediante polinomios locales (o *kernel*) estima una función de densidad de probabilidad, una función de regresión o sus derivadas mediante polinomios locales [24, 26]. Por último la tarea de la suavización Bayesiana de la serie de Fourier (o definida simplemente como *bayes*) es regresar una función cosenoidal optimizando el número de coeficientes y los valores que estos toman. La función 3.2 representa la serie de Fourier a suavizar mediante la teoria bayesiana.

$$f(t) = \sum_{l=1}^{L} al \cos\left(\frac{2\pi t}{T}\right) \tag{3.2}$$

Aplicando la función 3.3.

$$P(a_l) \rightarrow P(f_t) \Rightarrow < f >= \int f(t)P(f_t)df_t \tag{3.3}$$

Y calculando el error de estimación como 3.4.

$$\sigma_t^2 =< f^2 > - < f >^2 \tag{3.4}$$

Teniendo en cuenta 3.5.

$$P(\vec{a}, L|M) = \frac{1}{z}q(L, \vec{a})h(\mu|\vec{a}, L) \tag{3.5}$$

Donde $h(\mu|\vec{a}, L)$ representa la función de error como se muestra en 3.6.

$$h(\mu|\vec{a}, L) = \prod_{m=1}^{\mu} N[f(\vec{a}, L), \sigma_m^2] \tag{3.6}$$

Aplicando logaritmo natural a P se obtiene 3.7. Donde los hiperparámetros son definidos por $\vec{\theta}$.

$$\ln P = \ln q + \ln h - \ln z = F(\vec{a}, L, \vec{\theta}) \tag{3.7}$$

Tratando de maximizar una estimación *a posteriori* de $\vec{a}, L, \vec{\theta}$ que maximice $F(\vec{a}, L)$ suponemos que $\ln h$ es 3.8.

$$\ln h = \sum_{m=1}^{M} \ln N_m = -\sum_{m=1}^{M} \frac{(\hat{y_m} - f(\vec{a}, L))^2}{\sigma_m^2} - \frac{\mu}{2}(2\pi\sigma_m^2) \quad ; \sigma_m^2 = \sigma^2 \tag{3.8}$$

Por lo que $\ln P$ queda de la manera que se muestra en 3.9.

$$\ln P = \ln q - \frac{1}{\sigma^2}\sum_{m=1}^{M}[\hat{y_m} - f(\vec{a}, L)]^2 - \frac{\mu}{2}(2\pi\sigma_m^2) \tag{3.9}$$

Si se supone $q(L, \vec{a})$ como en 3.10.

$$q(L, \vec{a}) = q(L)q(\vec{a}) = q(L) \cong N(< L >; \sigma_L^2) \tag{3.10}$$

Por lo que al final se llega a la ecuación 3.11

$$\ln P = -\frac{1}{\sigma_L^2}[L - < L >]^2 - \frac{1}{2}(2\pi\sigma_L^2) - \frac{1}{\sigma^2}\sum_{m=1}^{M}[\hat{y_m} - f(\vec{a}, L)]^2 - \frac{\mu}{2}(2\pi\sigma^2) \tag{3.11}$$

Donde $f(\vec{a}, L)]^2$ representa la función de la serie de Fourier a optimizar.

El método implementado en realidad trata de una aproximación al esquema Bayesiano completo, es decir, la distribución *a posteriori* se aproxima obteniendo el máximo de dicha distribución. A ésta estrategia se le llama estimación máxima *a posteriori* (MAP, por sus siglas en inglés). Lo importante o ventajoso de los métodos de suavización que se proponenen en este texto es que evitan el extenso trabajo computacional de evaluar las funciones de *verosimilitud* en cada iteración de los modelos dinámicos. Ya que por ejemplo para evaluar una función de error $E(\vec{\theta})$, teniendo la función $f(\vec{x_t}, \vec{\theta})$ y además $\dot{\vec{x}} = g(\dot{x}, \dot{\theta})$ se tendría que evaluar cada una de las soluciones mediante una técnica recurrente, siendo la más conocida la red neuronal recurrente [27]. Por lo anterior el presente estudio propone aplicar métodos mucho más genéricos dependientes de los los datos disponibles y las interacciones de interés con el fin de ser más efectivos con menores tiempos y esfuerzos computacionales.

3.4 Inferencia de parámetros

Como se ha descrito anteriormente el objetivo del estudio es obtener la forma más probable de cómo estan interactuando los genes, es decir como se comporta el sistema biológico estudiado y como se conecta la red de regulación genética. Con ayuda de lo anterior se puede llegar a obtener los parámetros de dicha red con el fin

de ayudar a reproducir los resultados de los experimentos desarrolados por expertos en procesos biológicos.

Para optimizar los parámentros se utilizó el paquete *optim* que incluye R [24] minimizando el error cuadrado medio mostrado en 3.12, lo cual corresponde de igual manera a maximizar la *verosimilitud*.

$$ECM = \frac{1}{\mu}\sum_{m=1}^{M}[(y_{m_t} - f_{\vec{a}}(t))^2] \tag{3.12}$$

Capítulo 4

Resultados

En el presente capítulo se muestran los resultados obtenidos en base a la metodología propuesta y experimentación realizada en los capítulos anteriores. En primer lugar se muestran las simulaciones realizadas con los modelos empleados. Enseguida se exponene los resultados obtenidos con los muestreos previamente descritos. Después se presentan los resultados alcanzados con los modelos de suavización de datos implementados y por último se describe la inferencia de paramétros realizada a las redes estudiadas.

4.1 Simulación

En el presente trabajo se dió a la tarea de simular la red de interacción genética descrita anteriormente conocida en la literatura como *Represilador* utilizando el software libre R [24]. Con lo que se obtubieron las Figuras 4.1 y 4.2 para el caso de 3 genes. En ellas se observa la concentración de proteínas y la concentración de RNAmensajero de cada uno de los genes involucrados respectivamente, las cuales muestran el comportamiento oscilatorio descrito anteriormente. Éste último consiste en que mientras la concentración de un gen aumenta, la concentración del otro se ve influenciada negativamente por la relación que exite entre ambos. Este proceso se observa en las dos gráficas, tanto la concentración de proteínas como la concentración de RNAmensajero.

Después se realizó la simulación del *Represilador Generalizado* con el cual se

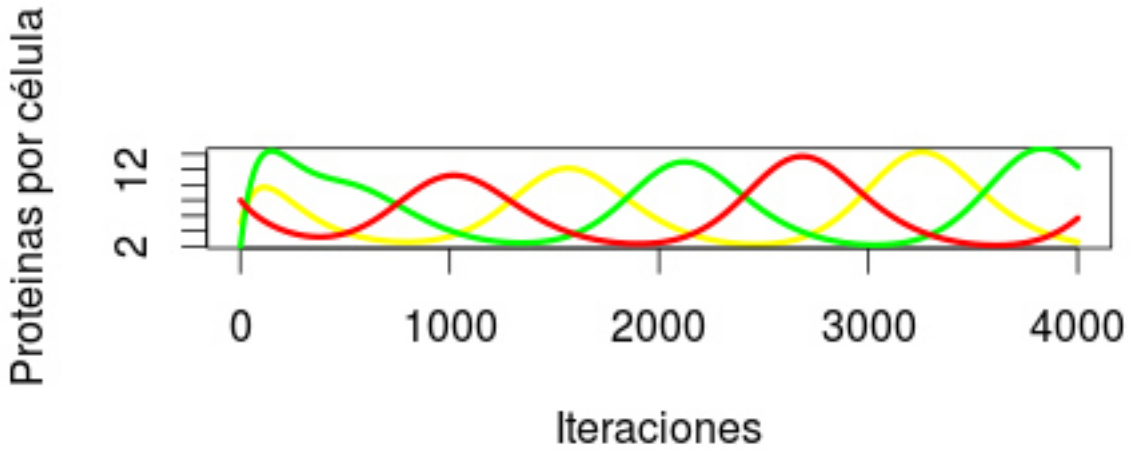


Figura 4.1: Simulación de concentración de proteínas por célula

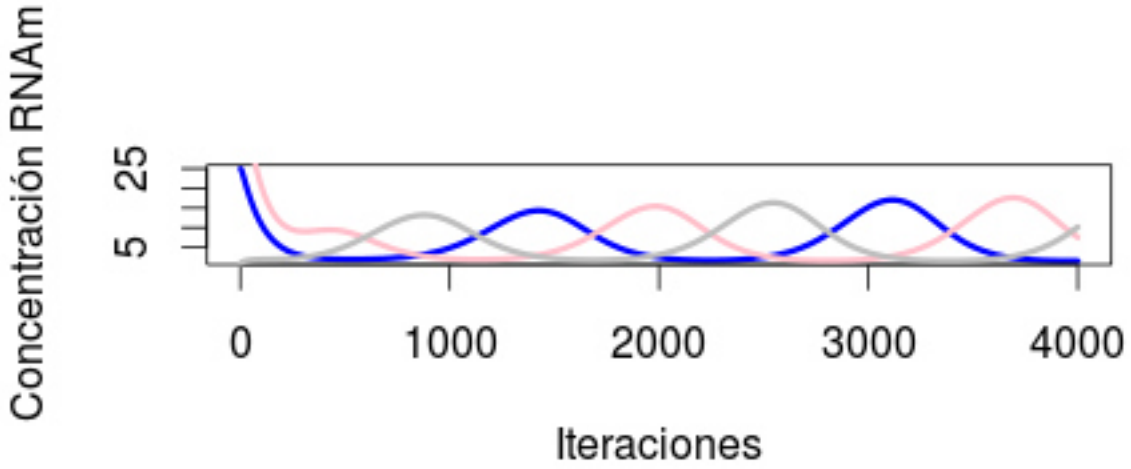


Figura 4.2: Simulación de concentración de RNAmensajero

puede reproducir la red de regulación genética con n genes. Es importante mencionar que entre las características inmersas en este sistema encontramos una sumamente interensante, la cual consiste en que si n toma valores pares, la mitad de los genes convergen a una constante, mientras que los restantes convergen a cero. Por otro lado, si n es un número impar el sistema tendrá el comportamiento oscilatorio característico observado en las Figuras 4.1 y 4.2. Esto producido por la misma estructura que presenta dicha red.

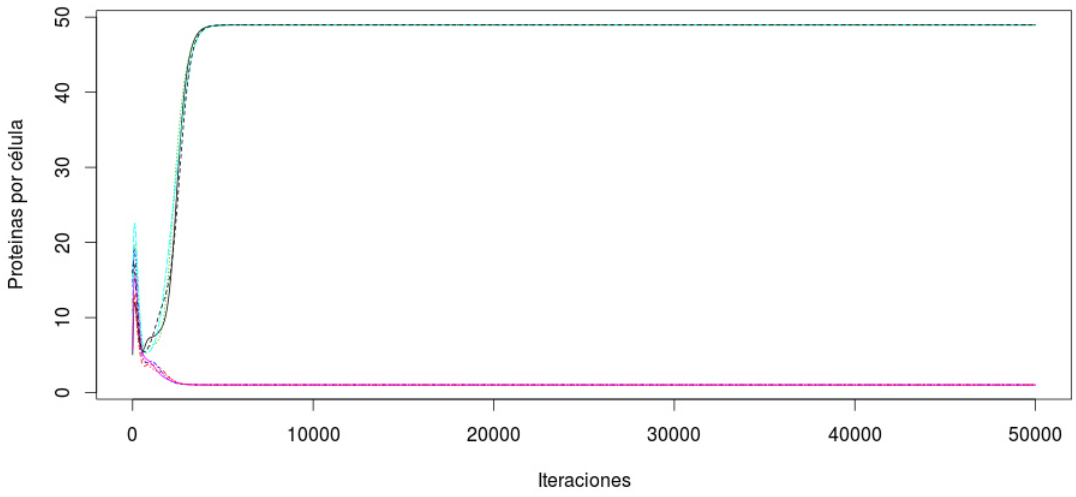


Figura 4.3: Simulación de concentración de proteínas por célula con 8 genes

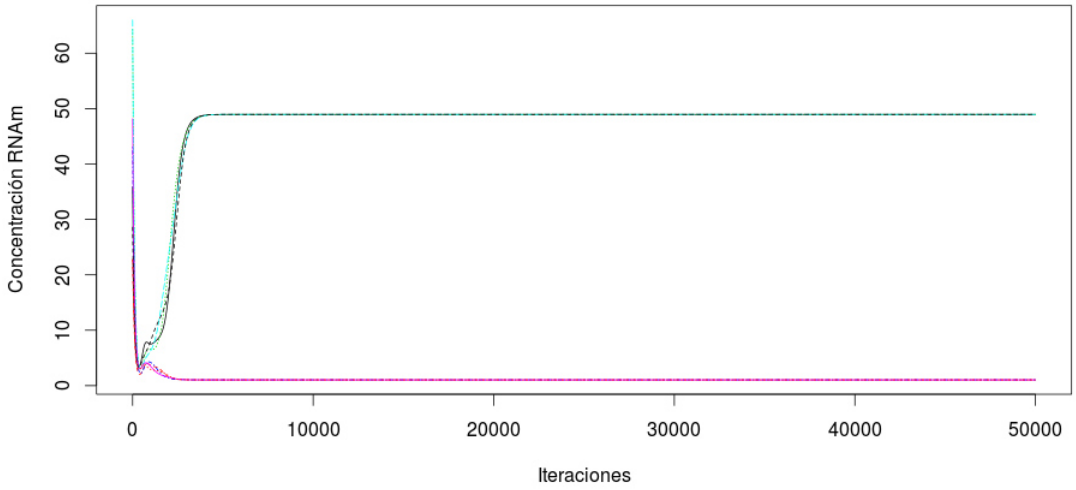


Figura 4.4: Simulación de concentracón de RNAmensajero con 8 genes

En la Figuras 4.3 y 4.4 observamos las simulaciones obtenidas en el caso de $n = 8$. Por otro lado podemos observar las Figuras 4.5 y 4.6 donde se muestran los casos de $n = 9$.

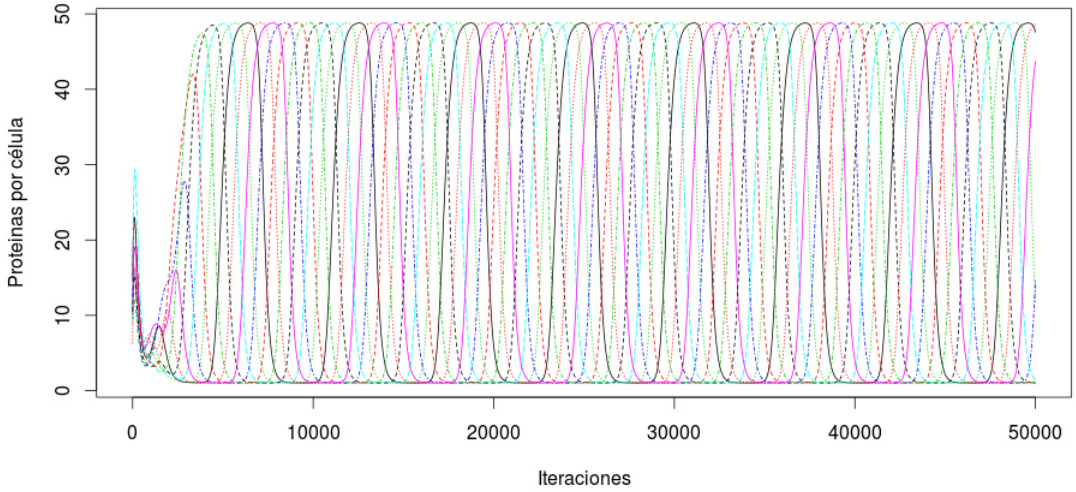


Figura 4.5: Simulación de concentración de proteínas por célula con 9 genes

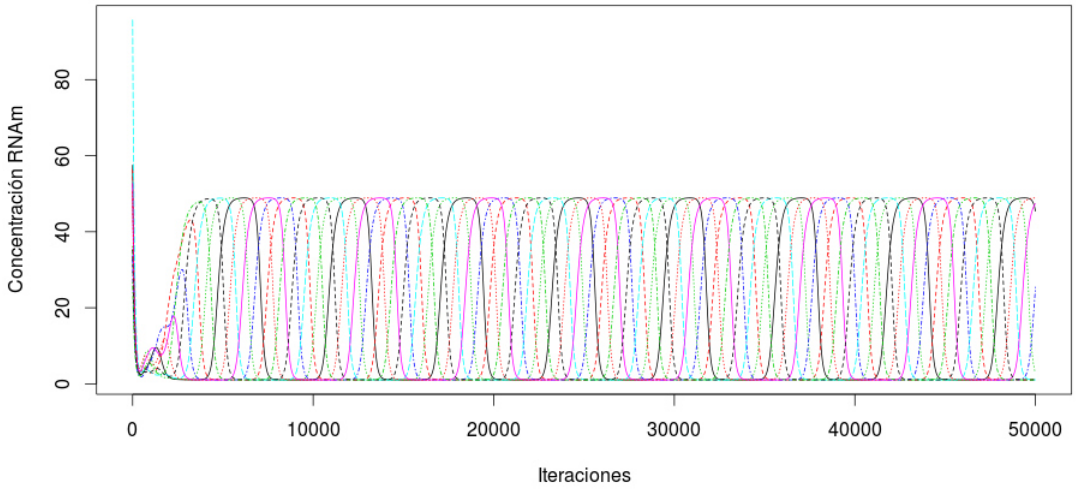


Figura 4.6: Simulación de concentracón de RNAmensajero con 9 genes

4.1.1 Métodos de muestreo

Como se descrbió en capítulos anteriores se emplearon dos tipos de muestreos; aleatorio simple y sistemático. Se realizaron ambos muestreos con diversos números de datos y niveles de error de medición relacionado con fallas humanas y perturbaciones en el entorno. Teniendo tres tamaños de datos y tres niveles de error como se muestra en la Tabla 4.1. Donde se hicieron las 9 combinaciones de la dimensión de los datos y el ruido presente en los puntos a la hora de muestrearse.

Tamaño de Muestra:	10	50	200
Nivel de Error:	1 %	5 %	10 %

Tabla 4.1: Descripción de Muestras

Por cuestiones de distinción solo se muestran las Figuras 4.7, 4.8 y 4.9. En éstas se observan los 3 tamaños de muestra con un error o falla de medición del 10 %, ya que se sabe que en procesos biológicos reales se tiene un ruido o error alrededor de esta magnitud.

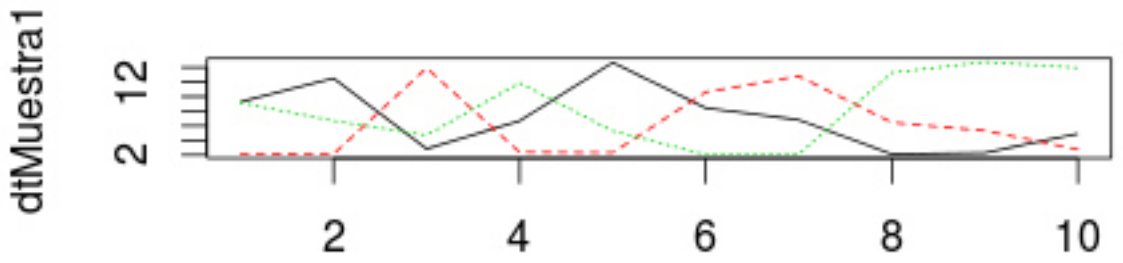


Figura 4.7: Muestra tamaño 10 con nivel de error de 10 %

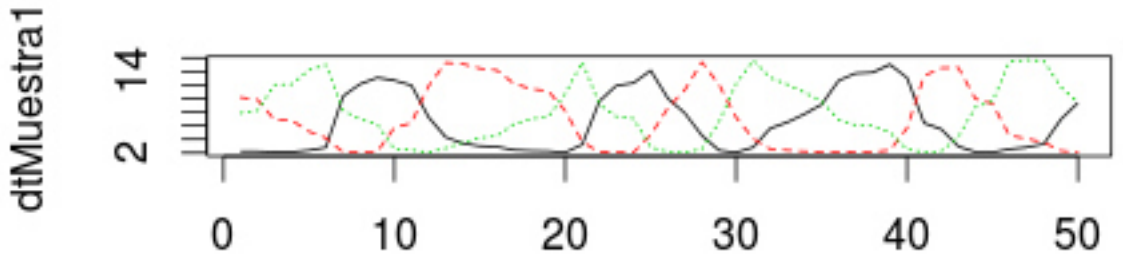


Figura 4.8: Muestra tamaño 50 con nivel de error de 10 %

En las tres Figuras se pueden contemplar los puntos graficados en base a los tres tamaños de muestra de la Tabla 4.1. Observando dichas gráficas se percibe acertadamente que mientras se incrementa el número de datos se obtiene una mejor

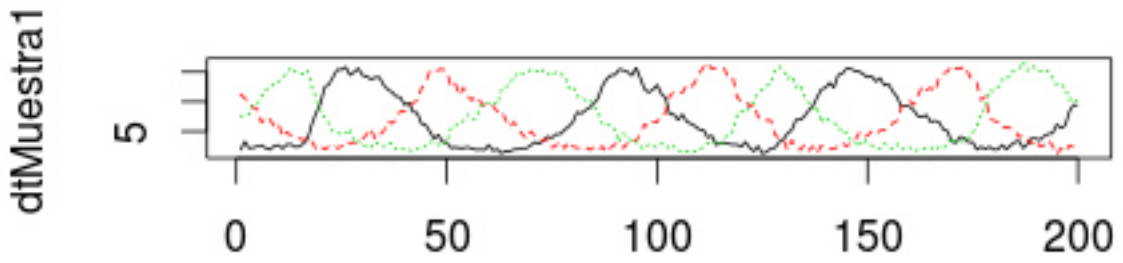


Figura 4.9: Muestra tamaño 200 con nivel de error de 10 %

aproximación al esquema original del sistema simulado. Por lo que se rescata el patrón oscilatorio del modelo aún teniendo un nivel de ruido grande (10 %).

4.2 Suavización

En esta sección se exponen los resultados obtenidos en base a los cuatro métodos de suavización que se experimentaron en el presente trabajo. Recordando que se tienen los métodos *spline*, *loess*, *kernel* y *bayes* descritos anteriormente en el Capitulo 3. Se diseñó una extensa experimentación donde se pretende explotar las características y propiedades de cada uno de los diferentes métodos planteados. En la Tabla 4.2 se resume los experimentos desarrolladas con cada uno de los procedimientos de suavización. En ésta se indica que se utilizan 10 diferentes tamaños de muestra que van desde 5 hasta 50 datos con incrementos de 5. Asimismo se tienen repeticiones temporales, estos puntos aluden a que en un experimento real se pueden repetir las mismas condiciones un determinado número de veces. Por último es importante mencionar que cada combinación del tamaño de muesta con las repeticiones temporales

se ejecutan 10 veces cada una.

Tamaño Muestra:	5	10	15	20	25	30	35	40	45	50
Repeticiones Temporales:	3		5		7		9		10	

Tabla 4.2: Diseño Experimental de Métodos de Suavización

En base al experimento se crea un total de 2000 gráficas. Ésto como resultado de los 10 duplicados de las combinaciones descritas en la Tabla 4.2 para cada uno de los cuatro métodos. Por cuestiones de distinción, ya que fueron los datos que se utilizaron en los análisis posteriores, se muestran los resultados de las gráficas de una tamaño de muestra de 20 y 3 repeticiones temporales, es decir, un total de 60 datos.

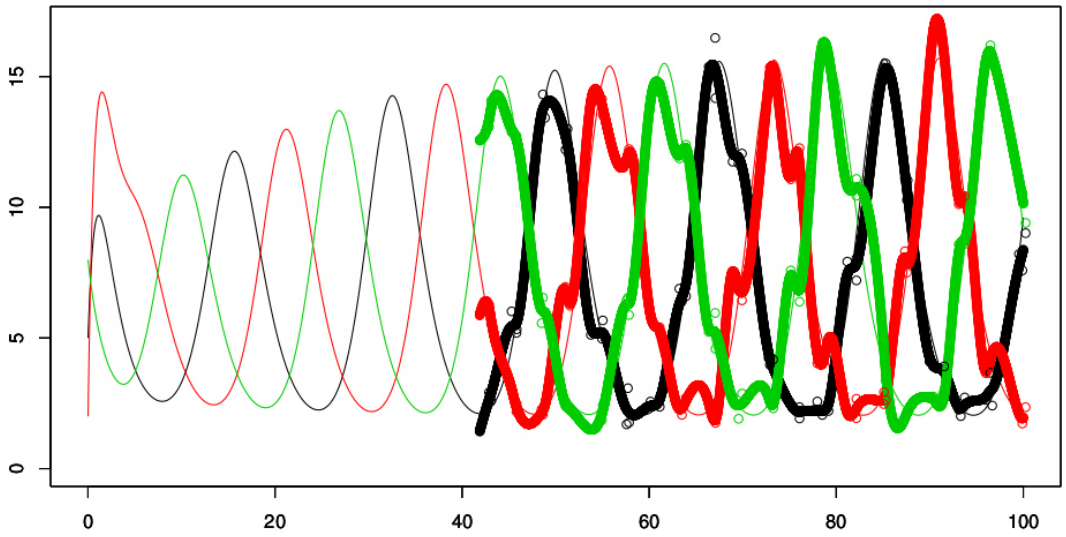


Figura 4.10: *spline* con tamaño 20 y 3 repeticones

En la Figura 4.10 se analiza el desempeño del método *spline* donde se muestra la simulación original en líneas finas y en líneas gruesas el suavizado logrado por

ésta función, donde el eje horizontal y el eje vertical representan el tiempo y la concentración de proteínas por gen respectivamente.

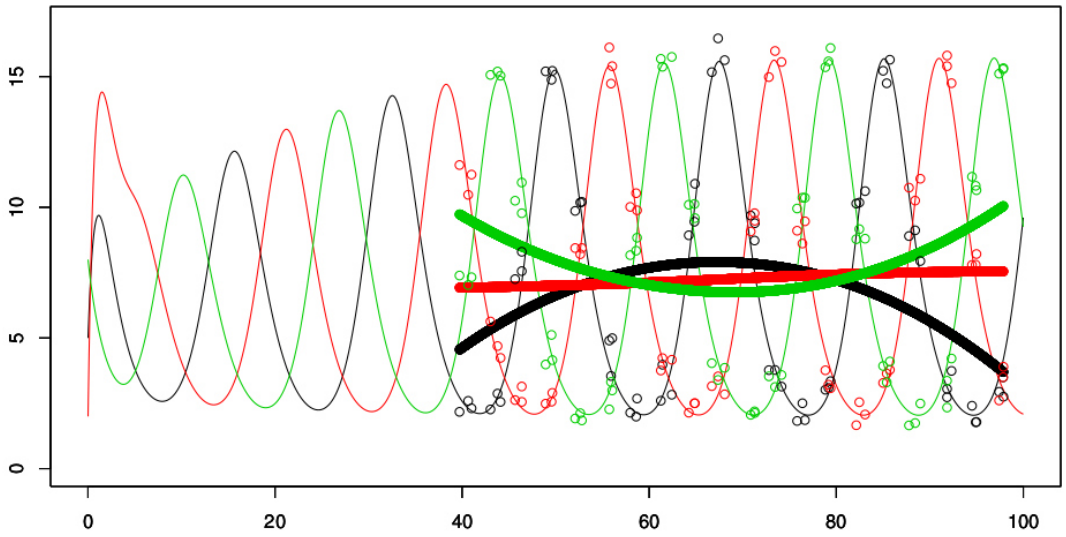


Figura 4.11: *loess* con tamaño 20 y 3 repeticones

Asimismo en la Figura 4.11 se muesta el suavizado alcanzado por del método *loess* donde de igual manera se muestra la simulación original en líneas finas y el suavizado en líneas gruesas.

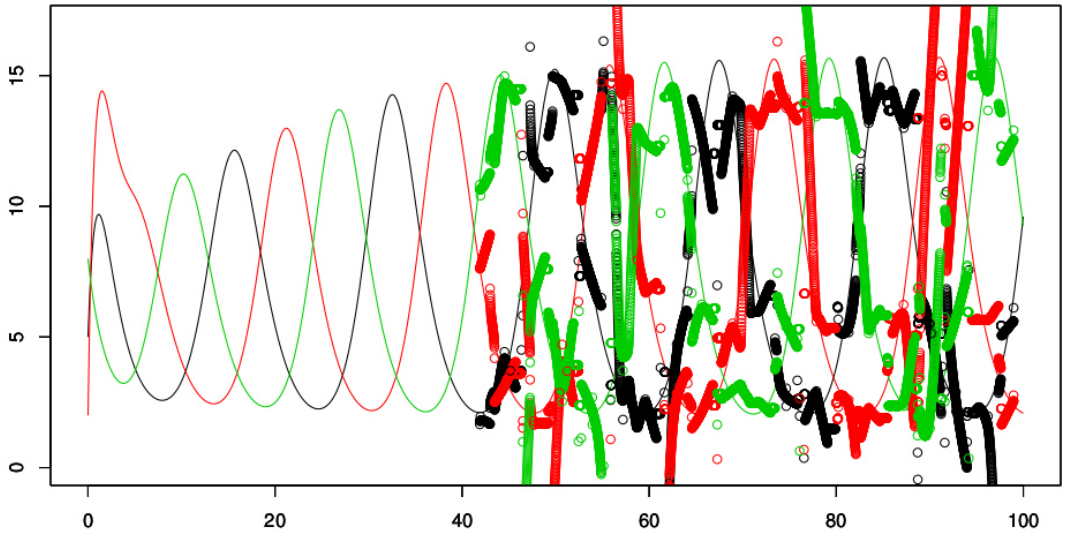


Figura 4.12: *kernel* con tamaño 20 y 3 repeticones

De igual manera en la Figura 4.12 se muesta la suavización lograda por el método *kernel* donde se muestra la simulación original en líneas finas y el suavizado en líneas gruesas.

Por último se muesta la Figura 4.13 en la cual se observa el suavizado con líneas guresas y la simulación original con líneas finas.

Se puede contemplar de una manera deductiva la eficiencia y desempeño de los cuatro métodos. Donde por un lado vemos que las funciones *loess* y *kernel* no revelan una buena aproximación a los datos originales. Ya que éstas no detectan la naturaleza oscilatoria de los datos. Mientras que los métodos *spline* y *bayes*, si bien no replican completamente la estructura de los datos originales, recuperan muy bien la oscilaciones características del sistema.

Retomando los resultados obtenidos por los métodos *spline* y *bayes* se observa su buen desempeño ya que el primero rescata la estructura de los datos originales

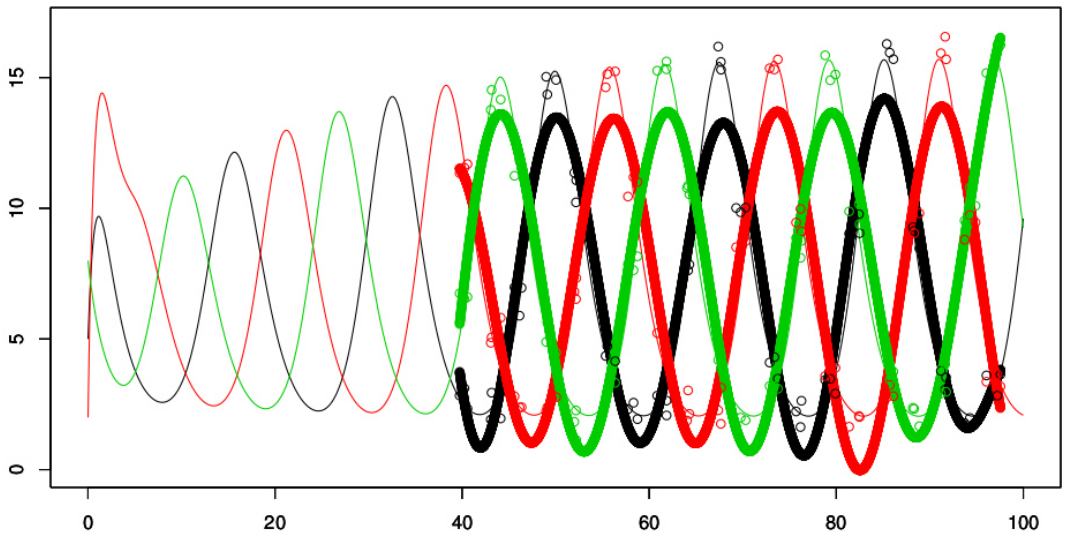


Figura 4.13: *bayes* con tamaño 20 y 3 repeticones

teniendo pequeñas variaciones alrededor de los puntos reales. Mientras que el segundo método recupera de forma adecuada la configuración de los datos reales solo con un pequeño defecto en las amplitudes. Pero en ambos casos se logra un objetivo importante al rescatar la peculiaridiad de la red genética estudida.

Es importante destacar el desempeño general de cada uno de los cuatro métodos basandose en el total de los experimentos realizados con todos los tamaños y repeticiones. Por esta razón se decidió hacer un gráfico que englobe todas estas pruebas. En la Figura 4.14 se muestra de manera descriptiva el rendimiento de las funciones de suavizado, donde se aprecia que el eje horizontal representa el número de puntos totales tomando en cuenta el tamaño de muestra como las repeticiones temporales. El eje vertical izquierdo muestra una error relativo a *chi* cuadrada, donde un error alrededor de 1 se considera aceptable, ya que uno mayor representa valores deficientes, mientras que con un error de 0 incurre a una sobreestimación de los datos. Finalmente en el eje vertical derecho se aclaran los colores y signos definidos para

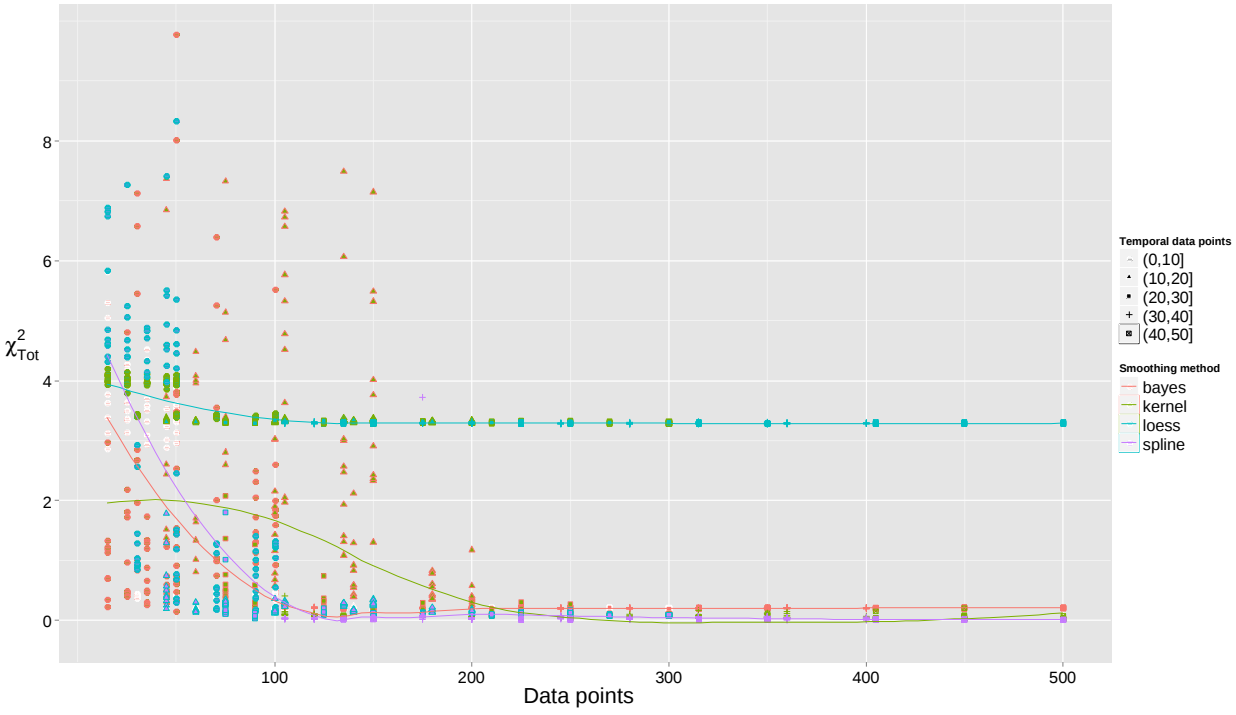


Figura 4.14: Comparación de Errores de Métodos de Suavización

cada función y cada caso respectivamente.

En esta Figura podemos observar que ambos métodos, tanto *spline* como *bayes* tienden a bajar su error más rápidamente que la función de *kernel*. Además se aprecia que el método *loess* nunca llega a ser tan efectivo aun teniendo una gran cantidad de datos de entrada. También es importante destacar que los dos primeros métodos llegan a trabajar considerablemente bien con menos de 100 datos, lo cual favorece a nuestra investigación ya que es de suma relevancia tener métodologias que sean robustas a la escasez de información.

Capítulo 5

Conclusiones

Finalmente este capítulo se discuten y analizan los resultados obtenidos en el presente trabajo. En la sección 5.1 se presentan las conclusiones generales de esta tesis, mientras que en la sección 5.2 se resumen las contribuciones aportadas por medio de ésta. Por ultimo en la sección 5.3 se detalla y propone el trabajo futuro a realizarse.

5.1 Conclusiones generales

En la presente investigación se estudió el desempeño y capacidad de las teorías bayesianas implementadas en metodologías de inferencia de parámetros en las que existen ocasiones en que los pocos y escasos datos representan un obstáculo en la obtención de estadísticos confiables y rápidos. Además se logró la implementación de las redes biológicas artificiales del *Represilador* y el *Represilador Generalizado* con el motivo de desarrollar y comporbar las metodologías del suavizado de los datos e inferencia de parámetros sobre éstos.

Se muestra el impacto que genera tanto la suavización bayesiana como la aproximación por regresión de polinomios locales en el importante ahorro de trabajo computacional y en la también relevante recuperación de las características y patrones. Aun así se destaca principalmente el método bayesiano ya que produce un error considerablemente menor trabajando con muestras pequeñas en comparación con la metodología de aproximación por regresión de polinomios locales.

5.2 Contribuciones

Las contribuciones derivadas del presente estudio se puntualizan de la siguiente manera:

- Se realizó la implementación efectiva de los modelos de redes genéticas estudiados como el *Represilador* y el *Represilador Generalizado*, asimismo se examinaron las características, bondades y propiedades que poseen.
- Se llevó a cabo la implementación de muestreo de datos para despúes utilizarlas en el análisis de las metodologías de suavizado.
- Se implemento la incorporación del ruido (o error de medición) a las muestras obtenidas.
- Se desarrollo e implementó la suavización bayesiana de una serie de Fourier.
- Se emplearon las funciones *spline*, *loess* y *kernel* de R [24].
- Se proponenen las suavizaciones *spline* y *bayes* como métodos genéricos confiables que ahorran tiempo evitando los grandes esfuerzos y trabajos computacionales.
- Se obtuvó la inferencia de los parámetros mediante la optimización del error con la función *optim* en R [24].
- Se graficaron y analizaron los resultados tanto de las suavizaciones como de la inferencia de datos con el fin de razonar una nueva metodología con mayor rapidez y además que cuente con cierto nivel de acertividad en casos en los que se dispone de escasa información.

5.3 Trabajo Futuro

En el presente estudio se muestran grandes áreas de oportunidad en las que la investigación es sumamente extendible. Algunos de los casos que pueden derivarse de éste son los siguientes:

- Estudiar otros métodos de suavización de datos.
- Incorporar a la inferencia de parámetros otras funciones de optimización.
- Implementar modelos de redes genéticas que tengan interacciones tanto postivias como negativas entre sus elementos.
- Realizar estudios con diferentes funciones de error.
- Realizar experimentos utilizando diferentes funciones de series de Fourier.
- Extender a caso de estudio real en células de cáncer de mama.

Bibliografía

[1] Aguilar-Roblero, R., D. Granados-Fuentes, I. Caldelas, A. Salazar-Juárez y C. Escobar, «Bases neurales de la cronobiología humana: el sistema circadiano distribuido», *Golombek D, compilador. Cronobiología Humana; ritmos y relojes biológicos en la salud y en la enfermedad. Buenos Aires, Argentina: Universidad Nacional de Quilmes Ediciones*, págs. 67–83, 2002.

[2] Aloy, P. y R. B. Russell, «Structure-based systems biology: a zoom lens for the cell», *FEBS letters*, **579**(8), págs. 1854–1858, 2005.

[3] Bertsch, S. M., «La teoria que nunca murió», , 2012.

[4] Borodina, I. y J. Nielsen, «From genomes to in silico cells via metabolic networks», *Current Opinion in Biotechnology*, **16**(3), págs. 350–355, 2005.

[5] Buse, O., R. Pérez y A. Kuznetsov, «Dynamical properties of the repressilator model», *Physical Review E*, **81**(6), pág. 066 206, 2010.

[6] Cochram, W., «Técnicas de muestreo», *Cecsa, México*, 1980.

[7] Davidson, E. H. y D. H. Erwin, «Gene regulatory networks and the evolution of animal body plans», *Science*, **311**(5762), págs. 796–800, 2006.

[8] Devore, J., *Probabilidad Y Estadistica Para Ingenieria Y Ciencias/Probability And Statistics For Engineering And Sciences*, Cengage Learning Editores, 2008.

[9] DEVORE, J., *Probability and Statistics for Engineering and the Sciences*, Cengage Learning, 2011.

[10] ELOWITZ, M. B. y S. LEIBLER, «A synthetic oscillatory network of transcriptional regulators», *Nature*, **403**(6767), págs. 335–338, 2000.

[11] GOLDSMITH JR, T. T. y M. E. RAY, «Cathode-ray tube amusement device», US Patent 2,455,992, diciembre 14 1948.

[12] HORI, Y. y S. HARA, «Oscillation pattern analysis for gene regulatory networks with negative cyclic feedback», en *Decision and Control (CDC), 2010 49th IEEE Conference on*, IEEE, págs. 5798–5803, 2010.

[13] HORI, Y., T.-H. KIM y S. HARA, «Existence criteria of periodic oscillations in cyclic gene regulatory networks», *Automatica*, **47**(6), págs. 1203–1209, 2011.

[14] LEE, T. I., N. J. RINALDI, F. ROBERT, D. T. ODOM, Z. BAR-JOSEPH, G. K. GERBER, N. M. HANNETT, C. T. HARBISON, C. M. THOMPSON, I. SIMON *et al.*, «Transcriptional regulatory networks in Saccharomyces cerevisiae», *Science*, **298**(5594), págs. 799–804, 2002.

[15] LOINGER, A. y O. BIHAM, «Stochastic simulations of the repressilator circuit», *Physical Review E*, **76**(5), pág. 051 917, 2007.

[16] LÓPEZ, M., G. RUIZ ROMERO y M. VEGA, «Biología de sistemas», *Genoma España/FUAM, Madrid*, 2007.

[17] MONTGOMERY, D. y G. RUNGER, *Probabilidad y Estadística Aplicadas a la Ingeniería [Applied Statistics and Probability for Engineers]*, McGraw-Hill, México, 1996.

[18] MÜLLER, S., J. HOFBAUER, L. ENDLER, C. FLAMM, S. WIDDER y P. SCHUSTER, «A generalized model of the repressilator», *Journal of Mathematical Biology*, **53**(6), págs. 905–937, 2006.

[19] SÁNCHEZ, J. M. C., *Inferencia estadística para economía y administración de empresas*, Centro de Estudios Ramón Areces, 1996.

[20] SHANNON, R. E., *Systems simulation: the art and science*, tomo 975, Prentice-Hall Englewood Cliffs, NJ, 1975.

[21] SIMONOFF, J. S., *Smoothing methods in statistics*, Springer, 1996.

[22] SINGLETON, P. *et al.*, *Bacteria in biology, biotechnology and medicine.*, Ed. 6, John Wiley & Sons, 2004.

[23] STRELKOWA, N. y M. BARAHONA, «Switchable genetic oscillator operating in quasi-stable mode», *Journal of the Royal Society Interface*, **7**(48), págs. 1071–1082, 2010.

[24] TEAM, R. C. *et al.*, «R: A language and environment for statistical computing», , 2005.

[25] WALPOLE, R., R. H. MYERS, S. L. MYERS y K. YE, «Probabilidad y estadística para ingeniería y ciencias», *Norma*, **162**, pág. 157, 2007.

[26] WASSERMAN, L., *All of nonparametric statistics*, tomo 4, Springer New York, 2006.

[27] WILLIAMS, R. J. y D. ZIPSER, «A learning algorithm for continually running fully recurrent neural networks», *Neural computation*, **1**(2), págs. 270–280, 1989.

[28] ZEISER, S., J. MÜLLER y V. LIEBSCHER, «Modeling the Hes1 oscillator», *Journal of Computational Biology*, **14**(7), págs. 984–1000, 2007.

[29] ZILL, D. G. y V. G. POZO, *Ecuaciones diferenciales con aplicaciones de modelado*, Thomson Learning, 2002.

Ficha autobiográfica

Brenda Aide Peña Cantu

Candidato para el grado de Maestro en Ciencias
en Ingeniería de Sistemas

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Tesis:

Inferencia de Parámetros en Líneas Celulares de Cáncer

Yo nací en la Ciudad de Monterrey, Nuevo León un Lunes 10 de Abril de 1989. Viví y crecí en la Ciudad de San Nicolás de los Garza, Nuevo Léon junto a mi madre Beatriz, mi padre Donaciano y mi harmano mayor Andrés. Fui alumna de la Preparatoria # 7 Oriente de la Universidad Autónoma de Nuevo León en el periodo 2004-2006. Doy gracias a mis padres por haberme dado la oportunidad de estudiar una carrera profesional, ya que ellos no corrieron con la misma suerte. Por lo que en el 2006 ingreso a la carrera de Ingeniero Mecánico Administrador dentro la Facultad de Ingeniería Mecánica y Eléctrica de la misma universidad. En el 2009 decidí seguir dentro de mis estudios la Orientación de Térmica y Fluidos. Asi mismo en Agosto del 2010 me integro al grupo de investigación de la *DOS* liderado por el Dr. Roger Z. Ríos. Terminé mi formación profesional el mes de Julio del 2011 realizando la

Tesis titulada *Evaluación de Métricas de Dispersión en Sistemas Territoriales* bajo la dirección del mismo Doctor en opción a obtener mi título profesional lograndolo el mes de Marzo del 2012 convirtiendome en la primera mujer ingeniera y la tercera en total de mi familia.

Posteriormente al haber concluido mi formación profesional, en Enero del 2012 me integro al Posgrado en Ingeniería de Sistemas como estudiante becario CONACyT de tiempo completo de maestría. Por lo que en Agosto del 2012 me sumo al grupo de investigación liderado por el Dr. Arturo Berrones donde comienzo a desarrollar este proyecto de investigación como requisito de titulación para obtener el grado de Maestro en Ciencias en Ingeniería de Sistemas.

Este es mi camino y esta es mi historia hasta el día de hoy...