Modelación Matemática de la Variabilidad Fenotípica del Sistema de regulación Lac en *Escherichia coli*

Yury Elena García Puerta

Maestría en Ciencias con Especialidad en Matemáticas Aplicadas

Centro de Investigación en Matemáticas (CIMAT)

Septiembre 2011

Co-asesor: Dr. Marcos Capistrán Ocampo

Co-asesor: Dr. Alexander de Luna Fors

"A veces sentimos que lo que hacemos es tan solo una gota en el mar, pero el mar sería menos si le faltara una gota"

madre teresa de Calcuta

Dedicatoria

"El destino une y separa a las personas, pero no existe ninguna fuerza que sea tan grande que haga olvidar a las personas que amamos"

Con todo mi amor ...

A mis padres Margarita y Albeiro, quienes me han brindado su apoyo incondicional en cada una de mis locuras.

A mi hermano Erwin,
a quien adoro con toda mi alma.

A mi no novio Jack,
quien es la fuerza de mi corazón.

Agradecimientos

Al Dr. Marcos Capistrán por su paciencia, dedicación y por sus aportes, no solo académicos, sino también a nivel personal.

Al Dr. Alexander de Luna, por darme la oportunidad de ser parte de este proyecto y por dejarme entrar a su laboratorio.

Al Dr. Luis Delaye por su colaboración en el análisis filogenético.

A Adrián Jinich, por su paciencia, disposición para trabajar conmigo, sus ideas y por todo lo que me enseñó.

A mis padres por su apoyo en cada cosa que se me ocurre hacer, especialmente a mi mamá por sus consejos y por creer tanto en mi.

A mis amigos Diego y Yuber, quienes me acompañarón en todo este proceso, gracias por escucharme y cuidarme cuando lo necesité. Gracias Diego, por creer en mi e invitarme a ser parte de este proyecto.

Al CIMAT por darme la oportunidad de disfrutar de sus beneficios y por todo lo vivido y aprendido en este lugar.

Al CONACyT por el apoyo económico durante estos dos años.

Y finalmente a Dios por permitirme hacer realidad mis sueños, por las oportunidades y las personas que a puesto en mi camino.

Resumen

Se quiere estudiar la variabilidad en los niveles de expresión del operón lac para una colección de 96 aislados naturales de *Escherichia* coli. El objetivo es establecer una relación de causalidad entre genotipo y fenotipo, es decir, cómo cambios en los genotipos producen cambios en los fenotipos del operón. Para esto se propone una estrategia que consiste en: 1) Caracterizar los fenotipos a través de un conjunto de parámetros, los cuales se obtienen del ajuste de un modelo matemático a datos obnetidos experimentalmente 2) Establecer relaciones entre las secuencias genéticas a partir de un análisis filogenético para finalmente relacionarlas con los conjuntos de parámetros obtenidos.

En este trabajo se presenta el modelo matemático, la solución del problema inverso para la recuperación de los parámetros y la propuesta de agrupamiento de las bacterias a partir de un análisis filogenético.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Planteamiento del Problema

Uno de los grandes retos de la biología moderna es entender los patrones de variabilidad entre individuos de una misma especia y sus posibles causas. Se sabe que la regulación de los genes, y no la diferencia en sus secuencias, es el factor principal que determina la diversidad entre los fenotipos de los individuos, esto es, sus características estructurales, fisiológicos o conductuales.

Las diferencias en la regulación de genes se deben tanto al medio ambiente que rodea a un individuo como a la "lógica" genéticamente codificada, que determina en qué circunstancias y a qué nivel de expresión se enciende un gen. [15, 7, 8, 28, 30, 39, 9, 16] Se conoce muy poco acerca del nivel de variabilidad natural de esta lógica de regulación genética y de las causas genotípicas y ambientales que subyacen estas diferencias.

Lograr entender este fenómeno es de suma importancia, pues hoy se sabe que

distintas enfermedades humanas de origen genético son causadas por cambios en la expresión genética como respuesta a ciertos efectos ambientales. Tal es el caso de la intolerancia a la lactosa, enfermedades cardiovasculares y algunos tipos de cáncer [19, 5, 6, 44, 45, 55].

En la actualidad existe poca información de la dinámica de este problema, dado que, la moyoría de los esfuerzos experimentales han estado enfocados en entender la variabilidad genética en humanos, con énfasis en su correlación con expresión genética en estados de salud y enfermedad. La dificultad con estos estudios es que los sistemas implicados son demasiado complejos para proporcionar un entendimiento profundo del problema de variabilidad natural en la regulació de expresión genética.

Sin embargo, en la naturaleza existen sistemas más simples y mejor caracterizados con los que se podría estudiar este fenómeno, como es el caso del operón lac. Dicha unidad será el objeto de estudio de este trabajo y el interés es proponer una estrategia que permita desarrollar un primer estudio de la variabilidad en los niveles de expresión del mismo en una colección de diferentes cepas de E.coli. Se quiere determinar si existe una relación de causalidad entre genotipos y fenotipos.

Una de las propuestas para resolver este problema consiste en caracterizar los fenotipos a través de un conjunto de parámetros correspondientes a constantes cinéticas de las reacciones involucradas en el funcionamiento del operón. Conocidos estos parámetros el paso a seguir es la implementación de estrategias de agrupamiento, tanto para éstos como para las secuencias, que permitan estable-

cer relaciones entre ambos.

En este trabajo se centra la atención en encontrar el modelo que describa "mejor" los datos y en solucionar el problema inverso, que consiste en recuperar los parámetros dados los datos, además se desarrollan algunas estrategias para la relación genotipo-fenotipo.

La colección de cepas para este estudio fue propocionada por la Dra Valeria Souza del instituto de ecología de la UNAM y en el laboratorio del Dr Alexander de Luna se secuenciarón un total de 1200 operones correspondientes a estas cepas de E.coli, de estas secuencias se hizo una preselección de 96, tomando aquellas con mutaciones; también se hicierón experimentos para medir los niveles de expresión del operón, en los que se evidencian variaciones; a estas variaciones se les conocerá como cambios fenotípicos.

El trabajo se divide en cuatro capítulos. El capítulo uno contextualiza el problema, dando un planteamiento del mismo y el objetivo que se quiere alcanzar.

En el capítulo dos se describe la dinámica del operón y se da una breve resumen de los trabajos de modelación que se han hecho del mismo. Para el análisis de los datos se tomarón modelos de la literatura, pero también se hizo un aporte de la deducción de un modelo, el cual es descrito en este capítulo; de igual forma se presentan los materiales y métodos utilizados para la selección del modelo, la identificabilidad de los parámetros, la solución del problema inverso y el trabajo con las secuencias genéticas.

En el capítulo tres se enseñan los resultados y finalmente en el capítulo cuatro se da la discusión y las conclusiones.

1.2. Antecedentes

El operón es una unidad genética de la E.coli involucrada en el metabolismo de azúcares, más adelante se explicará su mecanismo. Este sistema ha sido ampliamente estudiado. En 1960 Jacob et al dan el primer modelo biológico y poco tiempo después Goodwanes (1965) propone el primer modelo metemático [51]. Desde entonces se han propuesto modelos que describen su dinámica con diferentes niveles de detalle. En la literatura se encuentran tanto modelos deterministas como modelos estocásticos. En cuanto a este último, se sabe que la estocasticidad juega un papel importante en las fluctuaciones observadas en ciertos sistemas biológicos; la estocasticidad puede enriquecer el comportamiento del sistema, creando efectos que no son observados en el caso determinista [21]. Sin embargo, se han hecho estudios en los que se han relacionado las fluctuaciones fenotípicas con las respuestas biológicas a la adaptación, diferenciación y evoluación [47].

Autores como Yildirim, Ozbudack, Kaern, Santillán y Collins, entre otros han estudiado la dinámica del operón proponiendo modelos con diferentes niveles de detalle y con diferentes intereses de estudio, algunos se centran en estudiar el comportamiento cualitativo de esta unidad cuando se encuentra en medios tanto con glucosa como lactosa o inductores artificiales, otros se centran en el estudio de la estocasticidad que subyace al sistema o en el papel que juegan

mecanismos como la retroalimentación positiva y la negativa. En el capítulo dos se presenta un explicación de algunos de estos modelos.

Es de notar que, a pesar de todos los estudios que existen del operón, en la revisión bibliográfica que se hizo no se encontrarón experimentos que contaran con una variabilidad en las cepas, ni estudios que quisieran establecer relaciones entre genotipos y fenotipos.

1.3. Objetivo

Proponer un modelo matemático que de cuenta de la variabilidad fenotípica del operón, resolviendo el problema inverso para la estimación de los parámetros y dar estrategias para el análisis de las secuencias genéticas.

Capítulo 2

Materiales y Métodos

En este capítulo se describe la dinámica del operón además de los métodos y herramientas, tanto teóricas como numéricas utilizadas para la selección del modelo, la estimación de los parámetros y el análisis de las secuencias.

2.1. Descripción del Operón Lac

Las bacterias se alimentan de algunas fuentes de carbono como azúcares, preferiblemente glucosa, dado que ésta ingresa por difusión simple a través de la membrana, generando un ahorro energético para la célula.

En ausencia de glucosa, las bacterias se alimentan de azúcares más complejos tales como lactosa, maltosa, entre otros, pero esto requiere la expresión de diferentes genes para su metabolismo.

Cuando la bacteria *Escherichia* coli se encuentra en un medio rico en lactosa necesita de dos enzimas para poder metabolizarla; la *lactosa permeasa*, que

ayuda en el transporte al interior de la bacteria y la *β-Galactosidasa*, quien cataliza la hidrólisis de la lactosa en glucosa mas galactosa y derivados como la alolactosa.

Estas enzimas son sintetizadas por una unidad genética llamada operón (En el caso de la E.coli se conoce como operón lac)

El operón es una unidad genética, formada por un complejo de genes capaces de regular su propia expresión. En 1960 Jacob y Mondon [51] proponen un modelo genético del operón lac, que permite comprender cómo tiene lugar la regulación de la expresión génica en bacterias, este modelo sustenta que el operón consiste de tres genes estructurales, dos sitios de control y un gen regulador.

- **Gen Regulador (I)**: Secuencia de ADN que codifica para la proteína reguladora (represor) que reconoce la secuencia de la región del operador.

  Sitios de control

- **Operador (O)**: Región del ADN con una secuencia que es reconocida por la proteína reguladora (En el operón encontramos tres operadores, O_1, O_2 y O_3; uno antes del promotor, otro después del promotor y un tercero en la zona de lacZ).

- **Promotor (P)**: Región de ADN con una secuencia que es reconocida por la RNA polimerasa para comenzar la transcripción.

Genes estructurales

- **Gen Z**: Codifica para la *β-galactosidasa*, encargada del proceso de hidrólisis.
- **Gen Y**: Codifica para la *galactosido permeasa*, responsable del transporte.
- **Gen A**: Codifica para la *Tiogalactosido transacetilasa*, la cual cataliza la transferencia del grupo acetil del acetil Co-A al G-OH, de un aceptor tiogalatosido (este gen no está relacionado con el metabolismo de la lactosa.)

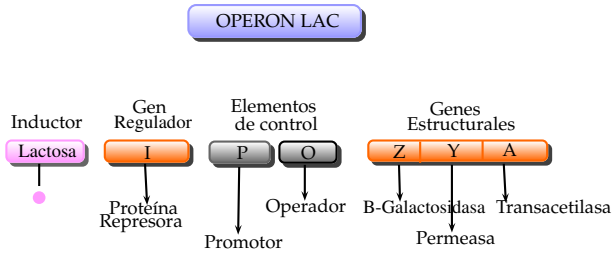


Figura 2.1: Estructura del Operón Lac

El operón puede encontrarse en dos estados, activo e inactivo, estos dependen de dos mecanismos de control, uno es el represor y el otro es el activador.

En ausencia de lactosa, el Gen I sintetiza una proteína llamada represor o inhibidor que se une al operador O, de modo que cuando la RNA polimerasa se une al promotor, no puede continuar la transcripción. (esto se conoce como control negativo-figura 2.2).

En presencia de lactosa, la alolactosa, un isómero de la lactosa, se une al represor cambiándolo conformacionalmente haciendo que no tenga afinidad por el operador (la alolactosa actúa como inductor), el operador queda libre y la

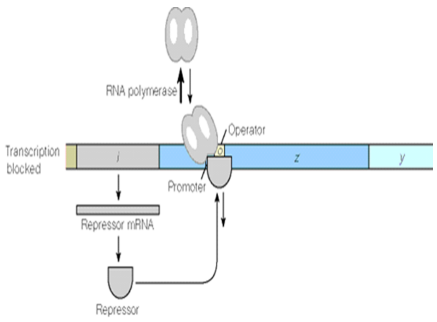


Figura 2.2: Acción del represor

RNA polimerasa puede realizar la transcripción, aumentando la síntesis de la β-Galactosidasa y la lactosa permeasa.

El mecanismo del operón está igualmente mediado por la cAMP y una proteína llamada receptora de cAMP (CAP: Proteína activadora catabólica). Esta proteína tiene sitios de unión para el ADN y cAMP, el complejo AMP-CAP se une a una zona cercana del promotor favoreciendo la unión de la RNA polimerasa, aumentando la transcripción de los genes (Este mecanismo se conoce como control positivo).

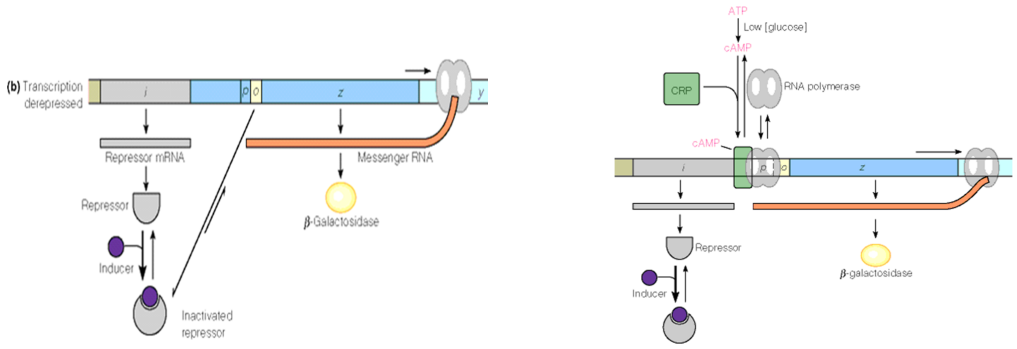


Figura 2.3: Acción del Inductor y el Complejo cAMP-CAP

Existen factores importantes en cuanto a la estructura de los represores, el represor es un tetrámero, es decir, está formado por cuatro subunidades, en donde cada subunidad puede unirse a una molécula de alolactosa; cuando las cuatro subunidades están libres el represor puede unirse a alguno de los operadores del operón e incluso puede unirse a dos operadores al mismo tiempo.

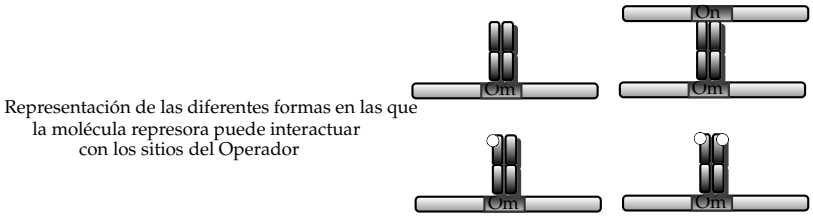


Figura 2.4: Interacción de los Represores y el Operador

Cuando dos de los sitios de unión del represor se encuentran ocupados por alolactosa, éste puede inactivarse o no dependiendo de la configuración que presente.

En el caso en que tres de los cuatro sitios de unión se encuentren ocupados por alolactosa, el represor queda totalmente inactivo.

Existen seis configuraciones diferentes que describen la interacción de un represor con dos moléculas de alolactosa.

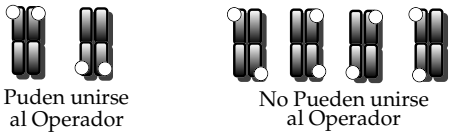


Figura 2.5: Interacción de los Represores y el Inductor

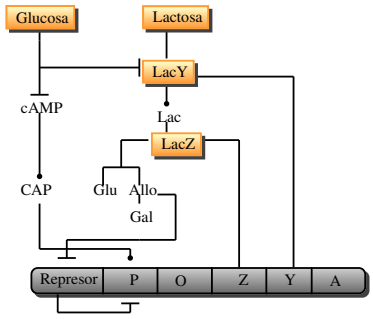


Figura 2.6: Principales Interacciones en el Operón

Sólo dos de las configuraciones permiten que el represor continúe activo. La figura 2.6 sintetiza las principales interacciones en el operón. Las terminaciones planas indican inhibición y las ovaladas, activación.

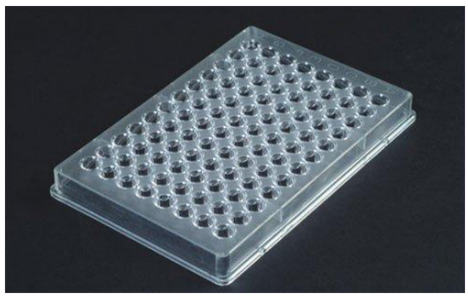
2.2. Experimentos

La cepa de E.coli usada para la generación de los datos fue $HG105$ y el plásmido $PUA66 - Lac1$. Inicialmente se contaba con una colección de aproximadamente 1200 cepas de E.coli aisladas de distintos húespedes, distintos ambientes y localizaciones geográficas. Dicha colección fue proporcionada por la Dra. Valeria Souza del Instituto de Ecología de la UNAM. En el laboratorio del Dr. Alexander de Luna Fors (Langebio-Cinvestav) se hizo una preselección de éstas, tomando aquellas con mutaciones en sus secuencias genéticas. En el conjunto de cepas seleccionadas se sustituyeron en el operón lac los genes estructurales β-galactosidasa, permeasa y transacetilasa por un gen reportero, que codifica para una proteína fluorescente verde (GFP). Se introdujo un número limitado de copias del plásmido, correspondientes a cada una de las 96 cepas del estudio, en células de E.coli HG105.

Para los experimentos de inducción de fluorescencia se crecieron las células en medio M9 a una temperatura de 35 grados centígrados en 12 concentraciones diferentes de inductor.

Las concentraciones utilizadas fueron {0, 0.98, 1.95, 3.9, 7.8, 15.625, 31.25, 62.5, 125, 250, 500, 1000}μM de IPTG. (La primera concentración corresponde a la concentración cero, la segunda a la concentración uno y así sucesivamente, cuando se haga referencia a la concentración cuatro se está hablando de la concentración que tiene como valor 7.8).

Las medidas registradas fueron nivel de fluorescencia y densidad óptica. Para los experimentos se usaron placas como las que se muestran en la gráfica.



Cada fila corresponde a una cepa y cada columna a una concentración de inductor, cada pozo contenía 150μl de medio y fue inoculado con 0,5μl de células.

Las lecturas se realizarón cada 45 min recogiendo un total de 24 lecturas, éstas se hacieron con la ayuda de una estación robotica ("EVO").

2.3. Proceso de Modelación

Para el proceso de modelación se hace uso de las leyes de la cinética química

Leyes de la Cinética Química

- Dada la reacción reversible $AB \longleftrightarrow A + B$ la velocidad de ambas combinaciones está dada por:

$$\frac{d[AB]}{dt} = k_{asoc}[A][B] - k_{dis}[AB]$$

- Para la reacción $AB \longleftrightarrow aA + bB$ la velocidad está dada por:

$$\frac{d[AB]}{dt} = k_{asoc}[A]^a[B]^b - k_{dis}[AB]$$

Este tipo de modelos son altamente complejos debido a:

- La cantidad de especies
- El número de reacciones involucradas
- La cantidad de constantes cinéticas
- Las constantes cinéticas tienen distintos órdenes de magnitud
- Estos modelos exhiben dinámicas en múltiples escalas de tiempo

Por lo que es necesario utilizar algunas estrategias para el manejo de las escalas de tiempo y poder simplificar la cantidad de ecuaciones resultantes. Algunos de estos resultados se presentan a continuación.

Método de reducción de modelos en sistemas de reacciones químicas [22]

Existen varios métodos para el manejo de las escalas de tiempo, entre ellos:

- lumping
- Análisis de sensibilidad
- Análisis en escala de tiempos
- Método escala en dos tiempos estándar
- Método de las variedades invariantes
- Método de la perturbación singular

Para los fines de este trabajo se usará el *Análisis en escala de tiempos* [22], con este tipo de análisis se supone que tras un periodo transitorio inicial, algunas reacciones rápidas pueden considerarse prácticamente instantáneas comparadas con las lentas.

Por lo que la ecuación diferencial que rige el comportamiento del sistema cambia de

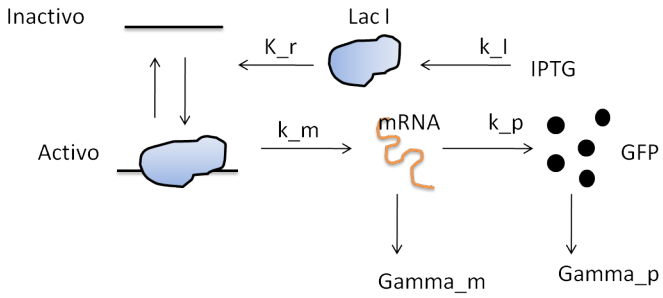
$$\frac{dc_s}{dt} = f_s(c_1, c_2...c_n)$$

A la ecuación:

$$f_s(c_1, c_2...c_n) = 0$$

Lo que nos da un conjunto de ecuaciones algebráicas que pueden ser acopladas con las ecuaciones diferenciales resultantes, esto nos ayuda a reducir el número de ecuaciones.

2.3.1. Deducción del Modelo Propuesto



Tipos de Reacciones

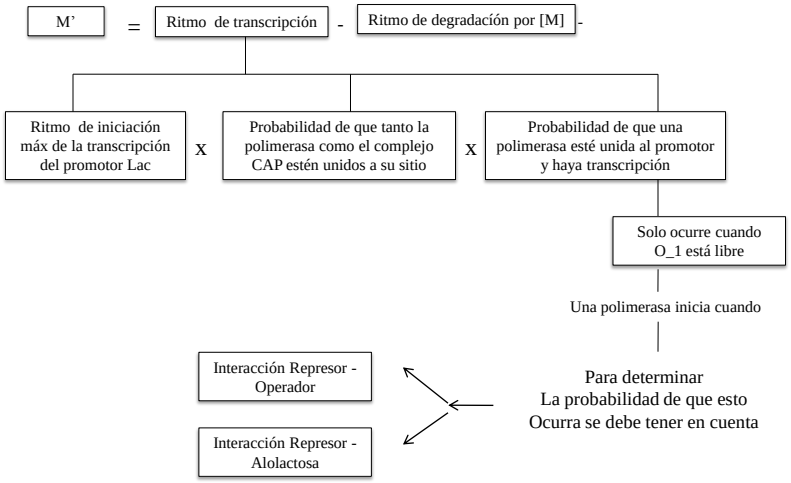
Proceso de transcripción (RNAm)	Unión polimerasa al promotor
Traducción de la proteína (GFP)	Interacción Represor-Inductor
Ingreso de IPTG a la célula	Interacción Represor-O_1
	Unión complejo CAP al promotor

Variables principales

- Concentración de mRNA [M]
- Concentración de la proteína (GFP) [P]

Dinámica del RNAm

El RNAm se produce vía transcripción del operón Lac y decrece por la actividad de degradación y debido al crecimiento celular.



Interacción Represor-Operador

Variable

- O: Operadores libres
- O_T: Total de operadores en la Bacteria
- O_R: Operador unido a un represor
- k_R: Constante de disociación represor-O_1
- R: Represor libre

Hipótesis

- Concentración total de operadores en la bacteria es constante (esto se debe a que estamos suponiendo que el número de copias de operones por bacteria es constante)

$$O_R + O = O_T \tag{2.1}$$

Reacción

$$R + O \overset{K_R}{\rightleftharpoons} O_R$$

Haciendo uso de las leyes de la cinética química:

$$\frac{dO}{dt} = k_1[O][R] - k_2[O_R] \tag{2.2}$$

Suponiendo un estado estacionario se tiene:

$$[R][O] = k_R[O_R] \tag{2.3}$$

donde $k_R = k_1/k_2$, despejando $[O_R]$ y sustituyendo en la ecuación de conservación se obtiene la probabilidad de tener el operador libre

$$\frac{[O]}{[O_T]} = \frac{k_R}{(k_R + R)} \tag{2.4}$$

Interacción Represor-IPTG

Variables

- R: Represor
- R_I: Represor unido a IPTG
- R_T: Total de represores
- k_I: Constante de disociación IPTG-Represor

- I: Concentración de IPTG

Hipótesis

Total de represores constante en la bacteria

$$[R] + [R_T] = R_I$$

Para esta reacción debemos tener en cuenta varios hechos importantes:

1. Las células inducidas y no inducidas crecen a un ritmo diferente, por tanto es importante considerar la dinámica de población a la hora de modelar [47].

2. *Como se explicó inicialmente, un represor no se inhibe al unirse a una sola molécula de IPTG debido a su estructura conformacional. Sabemos que se inhibe completamente al unirse con 3 moléculas de IPTG, pero siguiendo la propuesta de varios artículos [13,36,51] supondremos que esto se consigue sólo con 2 moléculas de IPTG para el caso del modelo 2 y supondremos que se inhibe con una sola para el modelo 4.*

Reacción

$$R + 2I \overset{K_I/V}{\rightleftharpoons} R_I$$

$$\frac{dR}{dt} = k_1[I][R] - k_2[R_I] \tag{2.5}$$

Igualando a cero:

$$[R][I] = \frac{k_I}{V}[R_I] \tag{2.6}$$

Despejando R_I y reemplazando en la ecuación de conservación se tiene la cantidad de represores libres:

$$[R] = \frac{(k_I V^{-1})R_T}{k_I V^{-1} + I^2} \tag{2.7}$$

Reemplazamos la ecuación 2.7 en 2.4

$$\frac{[O]}{[O_T]} = \frac{1 + k_I V^{-1} I^2}{(1 + k_R R_T + k_I V^{-1} I^2)} \tag{2.8}$$

Solo cuando O_1 está libre, puede una polimerasa unirse al promotor e iniciar la transcripción.

consideremos:

- $[N]$: Número de copias de Operon en el interior de la bacteria
- $[k_M]$: Máximo ritmo de transcripción del promotor
- γ_M: Degradación de mRNA
- M: Concentración de mRTNA

$$\frac{dM}{dt} = k_M[N]P_d prob(O) - \gamma_M[M] \tag{2.9}$$

Para los propósitos de este trabajo no incluiremos el término P_d, así:

$$\frac{dM}{dt} = k_M[N]\frac{1 + k_I V^{-1} I^2}{(1 + k_R R_T + k_I V^{-1} I^2)} - \gamma_M[M] \tag{2.10}$$

Dinámica de la Proteína (GFP)

Variables

- P : Proteína
- k_p: Ritmo de traducción de la mRNA
- γ_p: Degradación de la proteína

$$\frac{dP}{dt} = k_p[M] - \gamma_p[P] \tag{2.11}$$

Se supondrá, para los modelos 2, 3, y 4 que para el tiempo en que se inician las observaciones de la proteína las concentraciones de IPTG tanto a nivel extra como intracelular son iguales, por esta razón no se hace necesario proponer una ecuación para su ingreso.

La cuestión ahora es elegir una ecuación para el volumen que describa el fenómeno de manera apropiada. Algunos autores [13,14, 24, 51], han considerado el volumen celular en sus modelos y para este han hecho uso del cecimiento exponencial.

$$\upsilon = \frac{V}{V_0} = exp(\mu t)$$

Ajustaremos el modelo con este volumen (Modelo 3), pero también se ajustará con un crecimiento logístico (Modelo 2)

$$\upsilon = \frac{V}{V_0} = \frac{K}{exp(-\mu t)(K - V_0) + V_0}$$

Donde K es la capacidad de carga.

2.3.2. Deducción del Modelo-Munsky

Ingreso de IPTG

- El IPTG puede entrar y salir de la célula for difusión a través de la membrana
- El tiempo de degradación es muy grande
- El ingreso y salida de IPTG es proporcional al I_{out} e I_{in} respectivamente

$$\frac{dI}{dt} = r(I_{max} - I) \tag{2.12}$$

Producción de LacI

- Se produce de forma constitutiva
- Se pierde por degradación natural y cuando se inhibe con IPTG

$$\frac{dR}{dt} = k_l - (\delta^0 + \delta^1 I)R \tag{2.13}$$

Producción de GFP

- la producción de GFP depende de la probabilidad de que O_1 esté libre
- Los represores pueden inhibir la producción de GFP cuando dos de sus monómeros estén libre
- $O_R + O = O_T$

$$\frac{dP}{dt} = k_p prob(O) - \gamma_p P \tag{2.14}$$

Reacciones

$$2R + O \leftrightarrows OR$$

$$\frac{dR}{dt} = k_1[R]^2[O] - k_2[O_R]$$

$$k_1[R]^2[O] - k_2[OR] = 0$$

$$O_R = O_T - O$$

$$\frac{[O]}{O_T} = \frac{k_2}{k_2 + k_1[R]^2} = \frac{1}{1 + \alpha[R]^2}$$

2.3.3. Modelos

A continuación se presentan los modelos. Para los modelos 1 - 4 las variables principales son mRNA (M) y proteína (P) y para el modelo 5 son IPTG (I), mRNA del represor o lac I (R) y proteína (P).

Parámetros

Todas las medidas se llevarón a unidades de min y μM.

k_R	Constante de asociación represor-operon	$400x10^3$ $1/\mu M$	Stamatakis[43]
k_I	Velocidad de disociación IPTG-represor	$2{,}25x10^{-2}$ μM^{-2}	Yildirim[51]
k_M	Máximo ritmo de iniciación de la transcripción	$9{,}97x10^{-1}$ $\mu M/min$	Yildirim[51]
γ_M	Degradación mas dilusión de mRNA	$0{,}4111/min$	Yildirim[51]
k_p	Ritmo de traducción de la proteína	10 $1/min$	Yildirim[51]
γ_p	Degradación de la proteína	0,65 $1/min$	Yildirim[51]
μ	Crecimiento de la bacteria	$3{,}47x10^{-2}$ 1/min	Yildirim[51]
R_T	Total de represores	0,01 μM	Bionumbers
N	Número de plásmidos	2 mpb	Santillán [42]

Todos los parámetros fueron tomados de Munsky [35]

k_l	Producción de lacI	$1{,}7x10^{-3}$ 1/s	$1{,}2x10^{-1}$ 1/min
k_p	Producción de GFP	$1{,}0x10^{-1}$ 1/s	6.0 (1/min)
α	Fuerza de ocupación de lacI	$1{,}3x10^4$ N^{-n}	
γ_p	Degradación de GFP	$3{,}8x10^{-4}$ $N^{-1}s^{-1}$	$2{,}28x10^{-2}$ $N^{-1}min^{-1}$
δ^0	Degradación de Lac I	$3{,}1x10^{-4}$ $N^{-1}s^{-1}$	$1{,}86x10^{-2}$ $N^{-1}min^{-1}$
δ^1	Asociación lacI e IPTG	$5{,}0x10^{-2}$ $(\mu M.N)^{-1}s^{-1}$	3.0 $(muM.N)^{-1}min^{-1}$
r	Ingreso del IPTG al interior de la célula	$2{,}8x10^{-5}$ s^{-1}	$1{,}68x10^{-3}$ min^{-1}

1. **Modelo 1 - Guido**

$$\frac{dM}{dt} = k_M - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$

2. **Modelo 2**

$$\frac{dM}{dt} = k_M N \frac{1 + k_I V^{-1} I^2}{(1 + k_R R_T + k_I V^{-1} I^2)} - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$

Donde V corresponde a un crecimiento Logístico

$$V = \frac{\upsilon}{\upsilon_0} = \frac{K}{exp(-\mu t)(K - V_0) + V_0}$$

3. **Modelo 3**

$$\frac{dM}{dt} = k_M N \frac{1 + k_I V^{-1} I^2}{(1 + k_R R_T + k_I V^{-1} I^2)} - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$

V Corresponde a un crecimiento exponencial

$$V = \frac{\upsilon}{\upsilon_0} = exp(\mu t)$$

4. **Modelo 4**

$$\frac{dM}{dt} = k_M N \frac{1 + k_I V^{-1} I}{(1 + k_R R_T + k_I V^{-1} I)} - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$

Con crecimiento Logístico

$V = \frac{v}{v_0} = \frac{K}{exp(-\mu t)(K - V_0) + V_0}$

5. **Modelo 5-Munsky**

$$\frac{dI}{dt} = r(I_{max} - I)$$
$$\frac{dR}{dt} = k_l - (\delta^0 + \delta^1 I)R$$
$$\frac{dP}{dt} = \frac{k_p}{1 + \alpha R^2} - \gamma_p P$$

Acoplado con crecimiento Logístico $V = \frac{v}{v_0} = \frac{K}{exp(-\mu t)(K - V_0) + V_0}$

Es importante hacer notar algunos aspectos que jugarón un papel importante en la deducción de los modelos y su ajuste con los datos experimentales.

1. Los modelos propuestos describen el comportamiento del operón a nivel celular, mientras que los datos dan cuenta de una expresión a nivel poblacional, por lo que, es necesario incluir el crecimiento celular a la hora de comparar los resultados de los modelos con los datos expermentales, para esto se explorarón algunas alternativas, una de ellas consistió en incluir el crecimiento celular como un término más en los modelos, es el caso de los modelos 2, 3 y 4 (revisar la deducción para conocer las hipótesis al rescto). Otra estrategia utilizada consistió en multiplicar los resultados del modelo con la cantidad de bacterias en un tiempo dado a la hora de resolver el problema de optimización (mas adelante se presentarán los detalles), esta estrategia se utilizó con el modelo 5.

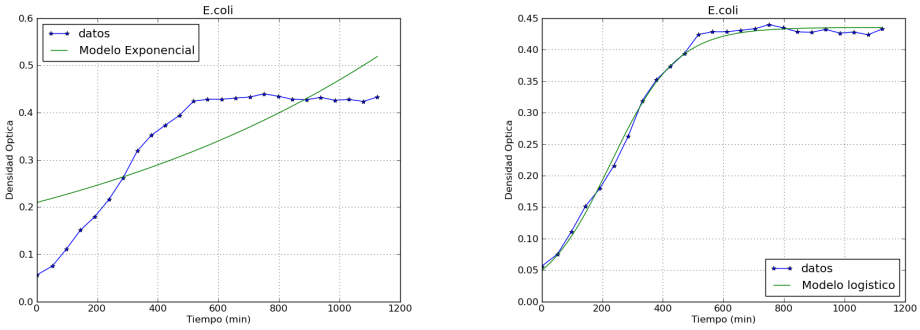


Figura 2.7: Problema Inverso

2. Qué tipo de crecimiento suponer para las bacterias?

   En la deducción de los modelos se mencionó que algunos autores lo ajustan con un crecimiento exponencial, pero en nuestro caso consideramos también un crecimiento logístico dado que describe mejor los datos de OD (ver figura).

   La densidad óptica (OD) corresponde a una medida de absorbancia (una medida de la luz transmitida a través de la suspensión). las suspensiones bacterianas dispersan la luz, al igual que cualquier partícula pequeña suspendida en agua. La dispersión de la luz es, dentro de ciertos límites, proporcional a la masa del cultivo.

3. Otro aspecto es el ingreso de IPTG a la bacteria, para los modelos 1-4 se consideró que los niveles intra y extracelular de IPTG eran iguales, mientras que, en el modelo 5 se da una ecuación extra para el ingreso de IPTG al interior de la bacteria.

   La diferencia entre los modelos dos y tres es la hipótesis del crecimiento celular, para el modelo dos se supone que sigue un crecimiento logístico,

mientras que, para el tres se supone un crecimiento exponencial.

Por otra parte, entre los modelos dos y cuatro la única diferencia es el exponente de I, para el modelo dos es 2 y para el cuantro es 1, esto tiene que ver con la conformación de los represores descrita en la dinámica del operón, el dos o el uno indican cuántas moléculas de alolactosa son necesarias para la inhibición de los represores.

2.4. Análisis de los modelos

Es de interés identificar un modelo que cumpla las siguientes condiciones:

- Que sea el modelo que explique mejor los datos usando criterios de teoría de la información.
- Que los parámetros del modelo correspondan a los componentes del operón, es decir, que tengan significado físico.
- Que el mapeo de parámetros a variables de estado sea inyectivo, esto es, que a cada conjunto de datos le corresponda un único conjunto de parámetros.
- Que los parámetros de interés sean identificables.

2.4.1. Selección del "Mejor"Modelo

El objetivo es elegir, de entre los cinco modelos, aquel que se ajuste mejor a los datos experimentales. Para ello se hace uso del criterio de información de Akaike (AIC)[23] y el criterio de información Bayesiana (BIC)[23].

Criterio de Información de Akaike(AIC)

Este criterio toma en cuenta que un modelo con más grados de libertad puede ajustar mejor los datos, en el sentido que la distancia entre los datos y los valores estimados es muy pequeña. Con este criterio se asigna a cada modelo un valor determinado por su máxima verosimilitud pero penalizado por el número de parámetros.

$$AIC = -2ln(\widehat{L}_i) + 2p_i$$

Donde $\widehat{L}_i$ denota la función de verosimilitud para el modelo i y p_i denota el número de parámetros de cada modelo.

Cuando se tienen pocos datos (n) en comparación con el número de parámetros, es decir, cuando $\frac{n}{q}$ pequeño (menor que 40), se hace uso de un criterio alternativo.

$$AIC_c = -2ln(\widehat{L}_i) + 2p_i \left(\frac{n}{n - p_i - 1} \right)$$

En este caso se tienen n = 24 mediciones y los modelos tienen 4, 10 y 13 parámetros, así que, se hará uso del criterio AIC_c.

La regla de decisión para la selección del modelo es escoger aquel con el más bajo AIC_c.

Criterio de Información Bayesiana (BIC)

Con este criterio se evalúa simultneamente la cualidad de cada modelo para explicar los datos. El BIC seleccionará el modelo entre las posibilidades que

mejor se aproxime al modelo real.

$$BIC = -2ln(\widehat{L}_i) + p_i ln(n)$$

n es el número total de tados usados en el proceso de estimación de parámetros.

2.4.2. Análisis de Sensibilidad

La eficiencia predictiva de los modelos es algunas veces obstaculizada por incertidumbre en los parámetros de entrada, dado que, en general, las ecuaciones que gobiernan o describen fenómenos dependen de varios parámetros.

El análisis de sensibilidad sirve para:

- Cuantificar la incertidumbre e identificar las relaciones entre los datos de entrada y de salida [23].
- El análisis de sensibilidad también cobra importancia cuando se quiere estudiar identificabilidad de los parámetros, dado que existen algunos métodos que hacen uso de la matriz de sensibilidad.
- Sirve para comprobar la lógica interna de un modelo, ayuda a entender como funciona el modelo o por qué no funciona correctamente.
- Para definir la importancia de cada parámetro, lo que servirá para determinar el grado de esfuerzo que debe prestarse a su medición de datos.
- Detectar si el modelo está sobreparametrizado, esto ocurre cuando existen parámetros a los que el modelo resulta insensible, en este caso será necesario eliminar algunos para simplificar el modelo.

Sensibilidad [23, 3]

Consideremos la ODE autónoma y no lineal

$$\frac{dy}{dt} = F(y; p)$$

Con condiciones iniciales $y_0 \in \mathbb{R}^m$ y parámetros $p \in \mathbb{R}^n$.

La solución de sensibilidad con respecto a los parámetros del modelo p_i, están definidos como el vector $s_i(t) = \frac{\partial y}{\partial p_i}$ y satisface la siguiente ecuación de sensibilidad.

$$\dot{s_i} = \frac{\partial F}{\partial y} s_i + \frac{\partial F}{\partial p_i}, \quad s_i(t_0) = \frac{\partial y_0(p)}{\partial p_i} \tag{2.15}$$

El objetivo es determinar el vector s_i, que corresponde a la derivada del sistema con respecto a cada uno de los parámetros.

2.4.3. Identificabilidad de Parámetros

Se puede hablar de Identificabilidad global e Identificabilidad local, a continuación se presentan las definiciones dadas por Miao et al [33].

Definición 2.4.1 *Globalmente Identificable*

El sistema dinámico dado por

$$\dot{x} = f(t, x(t), u(t), p)$$

$$y(t) = h(x(t), u(t), p)$$

Es globalmente identificable, si para unos datos de entrada dados $u(t)$ y algunos dos vectores de parámetros p_1 y p_2 en el espacio de parámetros P, $y(u, p_1) = y(u, p_2)$ si y sólo si $p_1 = p_2$

Definición 2.4.2 *Localmente Identificable*

El sistema dinámico dado por

$$\dot{x} = f(t, x(t), u(t), p)$$

$$y(t) = h(x(t), u(t), p)$$

Es localmente identificable, si para algún p en una vecindad abierta de algún punto $p*$ en el espacio de parámetros P, $y(u, p_1) = y(u, p_2)$ si y sólo si $p_1 = p_2$

Existe una amplia colección de teoremas y métodos [33, 38] para analizar la identificabilidad de los parámetros de un modelo.

Se pueden realizar varios tipos de sensibilidad, identificabilidad estructural, práctica e identificabilidad basada en sensibilidades. El análisis de identificabilidad estructural permite determinar de forma análitica si los parámetros son o no identificables dado un modelo; es un análisis que se realiza *a priori* al análisis numérico. En la mayoría de los casos se requiere de cálculos complejos. Este tipo de análisis permite decidir si es necesario remover o fijar algunos parámetros teniendo en cuenta las preguntas que se desean responder.

El análisis de identificabilidad práctica es empleado para la estimación de parámetros y la identificabilidad basada en sensibilidades es un análisis *a posteriori* en el cual, dado un punto del espacio de parámetros se puede determinar a partir de métodos numéricos si los parámetros son localmente identificables en un punto. En este trabajo sólo se hizo un análisis de identificabilidad estructural y práctico.

Para el análisis estructural, luego de una revisión bibliográfica [17, 33, 38] y algunos cálculos, se llegó a la conclusión que una expansión en series de Taylor

puede ser aplicado al modelo, dado que:

1. Es un teorema aplicable a modelos no lineales

2. Se puede determinar la identificabilidad de las condiciones iniciales.(Si se hace uso de otros teoremas solo es posible calcular ésta cuando las condiciones iniciales aparecen de manera explícita en el modelo). Este es un punto importante, dado que las condiciones iniciales también son parámetros en los modelos, aunque no se presenten de manera explícita.

3. Se puede realizar el análisis de identificabilidad a partir de la variable observable.

Teorema 2.4.1 ***Aproximación en series de Taylor***

Considere el sistema definido por:

$$\frac{dx(t)}{dt} = f(x(t), u(t), t, p)$$
$$x(0) = x_0(p)$$
$$y_m(t, p) = h(x(t), p)$$

Donde f y h son infinitamente continuamente diferenciable, sea:

$$a_k(\hat{p}) = \lim_{t \to 0^+} \frac{d^k}{dt^k} y_m(t, p)$$

$M(\hat{p}) = M(p)$ implica $a_k(\hat{p}) = a_k(p)$ para k= 0,1,2...

Una condición suficiente para que M sea identificable es por tanto $a_k(\hat{p}) = a_k(p)$ *para k= 0,1,2...*k_{max} *entonces* $\hat{p} = p$

la función $y_m(t,p) = h(x(t),p)$ hace referencia a la variable observable, en este caso la variable observable es P (Proteína - GFP) y M al sistema de ecuaciones.

En cuanto al análisis de identificabilidad práctica corresponde a la solución del problema inverso que se presenta a continuación.

2.5. Problema Inverso, Estimación de Parámetros

Es de interés, dado un conjunto de datos experimentales, recuperar o conocer aquellos parámetros que ajustan de manera óptima el modelo a los datos, este problema corresponde a la solución de un problema inverso.

Para la solución de este problema se tiene que la única variable de estado observable es la abundancia de GFP.

Denotemos por p el conjunto de parámetros que queremos identificar, y y^δ los datos con ruido, entonces el problema inverso se puede formular como un sistema de ecuaciones no lineales

$$F(p) = y^\delta \tag{2.16}$$

con

$$F(p) = \Psi(\Phi(p)), \quad \text{donde } F : \mathcal{P} \to \mathcal{D}$$

Mapea parámetros a datos.

$$\Phi : \mathcal{P} \to \mathcal{X}$$

Mapea parámetros a variables de estado.

$$\Psi : \mathcal{X} \to \mathcal{D}$$

Mapea variables de estado a datos.

Puesto que la ecuación 2.16 puede no tener solución debido a que el modelo es imperfecto y los datos tienen ruido, normalmente se resuelve el problema de optimización

$$\min_{p} ||F(p) - y^{\delta}||_Y^2 \tag{2.17}$$

Donde $|| \cdot ||_Y$ denota una norma apropiada para comparar la predicción del modelo y los datos. Para la elección de esa norma en nuestro caso supondremos que:

- Las mediciones de la abundancia de GFP en los cultivos son independientes.
- Siguen una distribución normal, con una desviación estándar constante σ.
- La media es la predicción del modelo de la dinámica de GFP.

Considerando estas hipótesis y un conjunto de datos de la forma

$$\{(t_1, {y_1}^{\delta}), (t_2, {y_2}^{\delta}), ..., (t_n, {y_n}^{\delta})\}$$

donde cada ${y_i}^\delta$ corresponde a las medidas de fluorescencia, se tiene:

$$Y_i^\delta \sim N(F(t_j, p), \sigma^2)$$

donde Y_i^δ denota la variable aleatoria de la observación en el tiempo t_i. Denotando por h_i las funciones de densidad correspondientes a las distribuciones Y_i^δ anteriores y bajo el supuesto que las observaciones son independientes entre sí, se puede expresar la densidad de la distribución conjunta F como:

$$F(y^\delta) = h_i({y_1}^\delta)h({y_2}^\delta)...h_n({y_n}^\delta)$$

la función de verosimilitud de la distribución Y^δ se puede expresar como:

$$L(p) \propto \prod_{j=1}^{N} h_i(y_i) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{F_j(p) - y_j^\delta}{2\sigma^2}) \tag{2.18}$$

Teniendo en cuenta lo anterior, definimos

$$||F(p) - y^\delta||_Y^2 = -\log(L(p))$$

$$\begin{aligned} log(L(p)) &\propto \sum_{j=1}^{n} \left[-\frac{1}{2}log(2\pi) - \frac{1}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(F_j(p) - y_j^\delta)^2 \right] \\ &= -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(F_j(p) - y_j^\delta)^2 \end{aligned}$$

Los dos primeros términos no dependen de p, luego, p es un máximo de $logL(p)$ si y sólo si p es un máximo de

$$-\sum_{i=1}^{n}(F_j(p) - y_j^\delta)^2$$

Si y sólo si es un mínimo de

$$\sum_{i=1}^{n}(F_j(p) - y_j^\delta)^2$$

Con lo que el problema se reduce a un problema de mínimos cuadrados.

Cuando se resuelve este problema se observa que aunque la distancia entre los resultados del modelo y los datos tiende a cero, los parámetros después de algunas iteraciones se alejan del vector de parámetros dado (Figura 2.8).

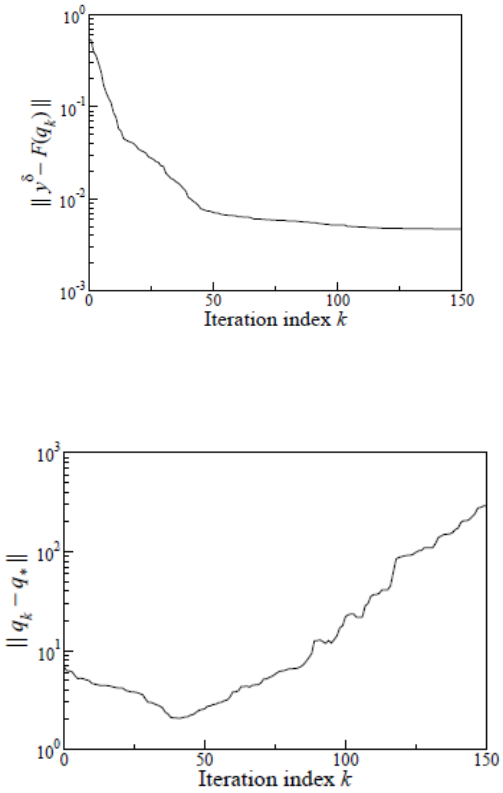


Figura 2.8: Problema Inverso

Esto se debe a que, en biología experimental, los datos experimentales tienen errores tanto técnicos como biológicos y estos se amplifican a la hora de hacer la estimación de parámetros [17]. Con el fin de resolver este problema e impon-

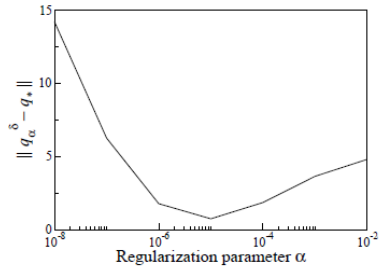
er condiciones adicionales para la identificación de los parámetros, se propone resolver el problema:

$$\min_{p} ||F(p) - datos||_Y^2 + \alpha R(p, p_0)$$

Donde $R(p, p_0)$ es un funcional de regularización.

Tomamos $R(p, p_0) = ||p - p_0||_1$. Este funcional de regularización se conoce como "sparsity enforcing regularization". Se sabe que si $p_0 = 0$, esta estrategia de regularización estabiliza la solución del problema inverso (minimiza la varianza) y obliga a la solución a que tenga un mínimo de entradas diferentes de cero [17].

El método de regularización muestra que existe un α para el cuál el error es mínimo



- Para α muy pequeño el error de los datos se amplifica
- Para α muy grande la solución se sesga

Por tanto, un problema a resolver es la elección de α. Se tiene que:

$$\alpha = \alpha(\delta)$$

Para dar un α óptimo es necesario tener información estadística de los datos y la única hipótesis que se está tomando es que cumplen una distribución normal, pero se desconoce información de la desviación estándar, por tanto, la estrategia que se propone es encontrar α con el algoritmo con el que se está haciendo la estimación de los parámetros.

Para la recuperación de los parámetros con el modelo cinco se explorarón tres estrategias:

La primera consiste en resolver, usando los datos correspondientes a una sola concentración de IPTG el siguiente problema:

$$\min_{p} ||F(p) * V(p) - GFP||_Y^2 + \alpha R(p, p_0)$$

Donde $V(p)$ corresponde a la cantidad de bacterias en un tiempo dado.

Por otra parte, se cuenta con mediciones de 12 concentraciones diferentes de inductor para cada cepa, es decir, se tienen 12 juegos de datos que corresponden al mismo fenómeno y por tanto deberían provenir del mismo modelo.

Con el fin de disminuir la varianza se hace una segunda propuesta, en la que se toman dos de esas concentraciones, en este caso, las concentraciones 4 y 5 que corresponden a niveles de inducción silvestre.

Es por ello que se propuso acoplar el modelo de Munsky con un crecimiento logístico para las bacterias.

Como se mostrará en los resultados, se explorarón algunas estrategias para la recuperación de los parámetros basadas en información a priori del fenómeno. Entre esta información se tiene que la densidad óptica del cultivo OD es proporcional a la cantidad de bacterias en un tiempo dado, $OD \approx \alpha V$, así que, puede intentarse recuperar los parámetros del volumen celular, no solo de las medidas de GFP, sino también de la densidad óptica, minimizando la siguiente distancia.

Sea ${Y_j}^{(i)}$ la observación de la i-ésima concentración (i=1,2) donde

$$ {Y_j}^i \sim N(F(t_j, p), \sigma^2) $$

Como los tiempos en los que se toman las mediciones son los mismos para las dos concentraciones, entonces, la distribución no depende de i, es decir, las dos observaciones en el tiempo t_j, provienen de variables aleatorias idénticamente distribuidas e independientes.

Por lo que, la función de verosimilitud de la distribución conjunta de cada tiempo, Y_j, será igual al producto de las verosimilitudes de cada ${Y_j}^i$, y a su vez, la función de verosimilitud de Y^δ es igual al producto de las verosimilitudes de cada Y_j, por lo tanto:

$$
\begin{aligned}
L(p) &\propto \prod_{j=1}^{n}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma}}exp((F_j(p)-{y_j}^i)^2 \\
&= \sum_{j=1}^{n}\sum_{i=1}^{2}\left[-\frac{1}{2}log(2\pi)-\frac{1}{2}log(\sigma^2)-\frac{1}{2\sigma^2}(F_j(p)-{y_j}^i)^2\right] \\
&= \sum_{i=1}^{2}\left[-\frac{n}{2}log(2\pi)-\frac{n}{2}log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{j=1}^{n}((F_j(p)-{y_j}^i)^2\right] \\
&= -nlog(2\pi)-nlog(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{2}\sum_{j=1}^{n}(F_j(p)-{y_j}^i)^2
\end{aligned}
$$

El problema se reduce a minimizar

$$\sum_{i=1}^{2}\sum_{j=1}^{n}(F_j(p)-{y_j}^i)^2$$

y finalmente, se usará este ajuste con dos concentraciones pero imponiendo una condición adicional para proponer una tercera estrategia. Se tiene que la desnsidad óptica del cultivo (OD) es proporcional a la cantidad de bacterias en un tiempo dato, $OD \approx \alpha V$, así que, puede intentarse recuperar los parámetros del volumen celular, no solo de las medidas de GFP, sino también de la densidad óptica, minizando la siguiente distancia

$$\min_{p} ||\beta V(p) - OD||_Y^2$$

2.6. Problema de Optimización

Para resolver el problema inverso se implementó un algoritmo genético (*Differential Evolution*), en lugar de un método de gradiente, teniendo en cuenta los resultados obtenidos por Moles and Mendel [26], quienes hicieron una

comparación de métodos de optimización en la solución de problemas inversos (Estimación de parámetros) para sistemas dinámicos no lineales de rutas bioquímicas, concluyendo que los métodos evolutivos ofrecen resultados más exitosos en la solución de este tipo de problemas.

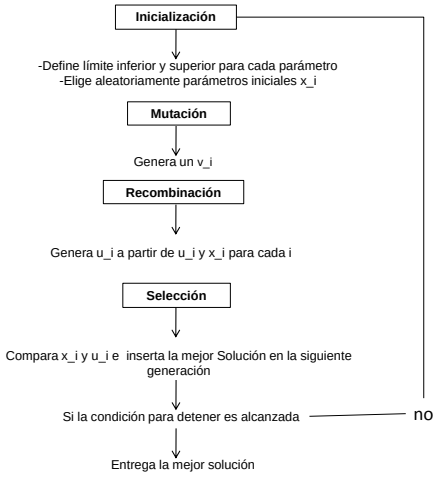
Un aspecto importante a definir con respecto a la *Differential Evolution* es una condición de paro. Zielinski, K. and Rainer L [55] hacen un estudio de diferentes criterios para detener la *Differential Evolution*, dichos criterios se pueden resumir en tres clases principales:

- Improvement-based criteria
- movement-based criteria
- distribution-based criteria

Luego de hacer pruebas con algunos modelos (alrededor de 15 modelos) cuya convergencia era conocida, encontrarón que la condición denominada como *Diff* ofrece un éxito del 100 % en todos los casos, excepto aquellos en los que la superficie o espacio de búsqueda es plano. Este criterio consiste en verificar si la diferencia entre el mejor y el peor valor de la función objetivo en cada generación se encuentra por debajo de un umbral dado por el usuario. Esta condición fue la implementada en el algoritmo (ED) junto con una condición adicional que corresponde a un número máximo de iteraciones.

Es preciso recordar que el problema inverso se debe resolver para 64 conjuntos de datos y el algoritmo tarda alrededor de 7 días en resolver el problema de optimización, por esta razón se hace necesaria la segunda condición de paro.

A continuación se presenta un esquema de la *Differential Evolution*



2.7. Mapeo Genotipo - Fenotipo

Se presentan dos posibles estrategias que pueden ser implementadas para el agrupamiento de las secuencias y el análisis entre el genotipo y el fenotipo.

2.7.1. Análisis Filogenético

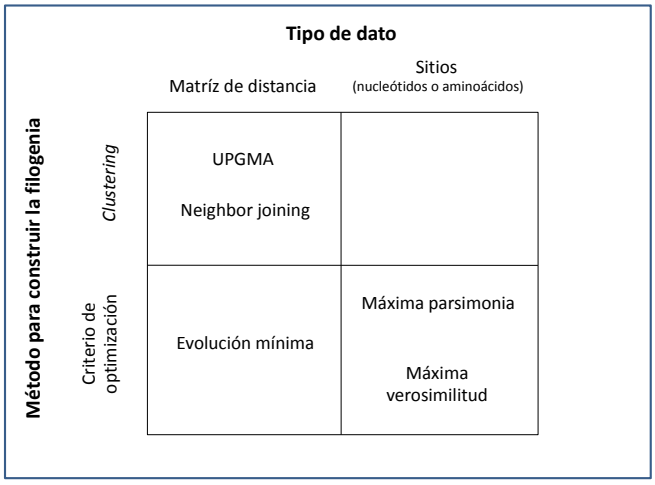
La primera propuesta es agrupar las secuencias mediante un análisis filogenético, el objetivo principal es reconstruir un esquema gráfico (árbol filogenético) que permita agrupar y clasificar evolutivamente las bacterias [1].

Los pasos a seguir para lograr este objetivo son:

- Escoger las OTUs o secuencias genéticas.

- Identificar las características que se tendrán en cuenta para agrupar.
- Analizar las características y formular una hipótesis sobre la homología que exista entre ellas.
- Construir una representación gráfica para la clasificación, esto es, representar la hipótesis formulada en el paso anterior en un esquema.

Existen diversas teorías para la evolución y de estas se desprenden diferentes estrategias para la contrucción de los árboles filogenéticos:



- **Neighbor Joining y UPGMA**: Son métodos basados en distancias. Estos métodos primero calculan la distancia total entre todas las parejas de taxones, teniendo en cuenta las diferencias en sus secuencias, y luego, calculan un árbol basado en esas distancias.
    - *Neighbor Joining*: Permite variación en las tasas evolutivas, el principio de este método es encontrar secuencialmente el vecino que minimiza la longitud total del árbol. Es un método robusto.

- *UPGMA*: No es un método muy recomendable para hacer análisis filogenético dado que considera la tasa de evolución constante y esto pocas veces es cierto

- **Máxima verosimilitud y Parsimonia**:Son métodos basados en caracteres, es decir, usan directamente las secuencias (secuencia de ADN o secuencia de proteína) de los taxones para determinar la relación con el ancestro más parecido. Son mejores filosóficamente dado que evalúan muchas hipótesis, pero son métodos muy costosos computacionalmente.

    - *Parsimonia*: Reconstruye un árbol filogenético (Cladograma) que representa la relación evolutiva de las especies en estudio. En la construcción del cladograma el método parsimonia busca la representación que tenga el menor número de cambios genéticos con respecto a la secuencia de un ancestro común. Evalúa todos los posibles árboles que representen la evolución y busca el más óptimo.

Para conseguir el árbol filogenético se usó MEGA, un programa para análisis filogenético y de acceso libre.

MEGA permite construir cada uno de estos árboles, así que, se hacieron los árboles para Neighbor Joining, Máxima Verosimilitud y Parsimonia, una vez se tengan estos árboles se comparan para determinar cuáles nodos se encuentran bien definidos (los nodos comunes a los tres árboles).

2.7.2. Correlación entre Distnacias

Una segunda propuesta consiste en establecer una correlación lineal entre las distancias de los parámetros y las distancias de las secuencias, la idea es implementar un álgoritmo que evalúe aleatoriamente regiones de la secuencia genética hasta determinar aquella que ofrece una mejor correlación entre las distancias. Una vez conseguido esto, podemos recuperar la región que mayor influencia tiene y analizar las mutaciones que en ella se presentan y a que región del operón pertenece.

Este tipo de estudios se puede hacer por regiones, rescatando aquellas de mayor interés, la región de LacI, la región promotora y la del operador y asociarlo sólo con las constantes cinéticas que guardan una estrecha relación con la región.

La idea original de este código es de Erez Dekel, asistente de investigación del Dr. Uri Alon en Weizmann Institute of Science.

2.7.3. Descripción de Algunos Modelos Reportados en la Literatura

Se identificarón en la literatura un conjunto de modelos que describen la dinámica del operón y unidades genéticas similares, estos son algunos:

Collins et al (2003) proponen una aproximación integrada, combinando teoría y experimento, para analizar y describir la dinámica de un módulo genético aislado.

Con su análisis logran establecer regiones de biestabilidad, un comportamiento muy común en estos sistemas. El mecanismo modelado no es propiamente el operón, sino que corresponde al fago-α, el cual cuenta con algunos mecanismos similares al del operón. Sin embargo, otros autores como Santillán(2007) y Yildirim(2003) modelan el operón y hacen una descripción cualitativa del mismo, demostrando que éste también muestra comportamientos de biestabilidad cuando la bacteria está en presencia de glucosa y lactosa al mismo tiempo.

Yildirim et al (2003) ofrecen un modelo del proceso de inducción en el operón lac incluyendo muchos detalles biólogicos, pero olvidando otros como la represión catábolica y detalles de la transcripción y translación. Su modelo consta de cinco ecuaciones diferenciales en donde las variables principales son la mRNA, la allolactosa, lactosa, β-Galactosidasa y la permeasa, sus ecuaciones son ecuaciones con retardo, es decir, tiene en cuenta los diferentes tiempos de los procesos. Como se mencionó arriba, Yildirim centra su estudio en un análisis cualitativo del modelo, determinando regiones de biestabilidad.

Ozbudack et al (2004) estudia la multiestabilidad del operón a partir de datos experimentales y un modelo muy sencillo del sistema, el cual consta de dos ecuaciones diferenciales, la primera da cuenta de la variabilidad de la permeasa (lacY) y la segunda de la captación del inductor; en su caso se trata de TMG. Un hecho importante es que su modelo no logra evidenciar la biestabilidad cuando se usa un inductor natural como la lactosa. Mas adelante, estos datos son usados por Santillán(2008) para la validación de su modelo y explica el por qué con los experimentos de Ozbudack no se logró biestabilidad cuando

se uso lactosa.

Kaern et al (2005) discuten los mecanismos teóricos que se creen causan las fluctuaciones en los niveles de expresión de un solo gen y los experimentos que han sido usados para validar esas ideas. De igual forma describe estudios experimentales de efectos estocásticos en redes de regulación genética. Propone un modelo estocástico y otro determinista.

Como resultados de su análisis obtiene:

- El tamaño del sistema y la cinética del promotor incrementan el ruido.
- La aproximación determinista no puede capturar los efectos potencialmente significantes que causan estocásticidad en la expresión de genes.
- Las condiciones necesarias para que las predicciones del modelo determinista y el estocástico sean similares son: un gran tamaño del sistema (Gran número de expresión de mRNA y proteína, y gran volumen celular) y una rápida cinética del promotor

Yusuke and Sano (2006) estudian el papel que juega la retroalimentación positiva y cómo los mecanismos de regulación permiten que la célula adopte múltiples estados de expresión internos. Aunque su modelo no corresponde exactamente al operón lac, si lo es de un mecanismo que tiene un comportamiento similar a este. El modelo no captura todo los mecanismos del sistema, sino aquellos que se consideran esenciales, es una versión simplificada del modelo propuesto por Collins (2003)

Guido et al(2006) analizan el comportamiento de redes genéticas sintéticas en condiciones cada vez más complejas: no regulada, reprimida, activada, y ambas, reprimida y activada. En principio proponen un modelo determinista y luego lo extienden, incluyéndo efectos estocásticos. Éste permite concluir que las fluctuaciones observadas en la actividad del promotor se deben al número de plásmidos, al crecimiento y división celular.

Kulman (2007) desarrolla un modelo termodinámico de regulación de genes en el que incluye los mecanismos conocidos de lacR-Inductor y el bucle de ADN y la cooperatividad de CRP. Se enfoca principalmente en dos aspectos:

- La interacción lacR-Inductor es solo dbilmente cooperativa. Para qué mecanismos la respuesta observada de inducción llega a ser abrupta?
- Es conocido que el CRP ayuda a mejorar la actividad transcripcional, pero al mismo tiempo facilita la formación de un bucle en el AND estabilizando la unión entre lacR y el DNA. La pregunta a contestar es: Qué rol funcional juega el CRP actualmente en el control del operón?

Santillán (2008) hace un estudio del comportamiento biestable del operón a partir de un modelo que consta de tres ecuaciones diferenciales, las variables principales son, RNAm, permeasa y lactosa. Desarrolla un modelo matemático muy completo en el que considera cada uno de los mecanismos de regulación conocidos para este sistema, estudia los efectos de inducción con lactosa y no con inductores artificiales, además considera algunos detalles tales como la interacción de los represores con los tres operadores, las diferentes configuraciones del represor y el inductor, la interacción del complejo CAP con el operador tres, el cual influye en el plegamiento del DNA, además considera el hecho que el ritmo

de crecimiento de la bacteria varía en concordancia con la captación de glucosa y metabolismo de la lactosa y estudia cómo esto afecta el comportamiento del sistema dinámico.

Munsky (2009) propone un modelo a partir de la ecuación maestra. Las variables principales son lacI y la proteína, donde la transcripción, translación y degradación son eventos que cambian los estados del sistema. Su principal objetivo es examinar la posibilidad de identificar los parámetros del sistema y mecanismos directamente de la distribución de células individuales, además de demostrar que el análisis de la variabilidad proporciona más información que solo el significado del comportamiento.

Stamatakis and Mantzaris (2009) derivan tanto un modelo determinista como uno estocástico para describir la dinámica del operón lac. Ambos modelos comparten detalles bioquímicos y se análisan a la par para demostrar efectos de estocásticidad. Tal comparación revela que en la presencia de estocásticidad, ciertos mecanismos bioquímicos pueden influir profundamente las regiones donde el sistema revela biestabilidad.

El modelo determinista se propone a partir de las leyes de la cinética química y se incluyen una gran cantidad de mecanismos conocidos sobre el operón, en tanto que el estocástico se hace a partir de la ecuación maestra.

Otros autores como Santillán (2010) y Hemberg (2007) han propuesto modelos estocásticos a partir de la ecuación maestra para estudiar la estocásticidad de sistemas de este tipo. Una de las principales diferencias en las propuestas de

estos modelos [21, 54, 37, 35, 43] es la elección de las funciones de propencidad y el tratamiento de las ecuaciones; algunos hacen uso de aproximaciones adiabáticas[54], otros expanden la ecuación en función de su tamaño [48] y otros se valen de simulaciones computacionales dado que hasta ahora no se conoce una forma de hacer inferencia con dicha ecuación.

Capítulo 3

Resultados

A continuación se presentan los resultados que se obtuvierón al analizar numérica y cualitativamente los modelos. En la primera parte se presentan los resultados que se obtienen al aplicar los criterios de teoría de la información a cada uno de los modelos.

Una vez realizada esta elección se procede a realizar un análisis cualitativo, un análisis de sensibilidad y uno de identificabilidad al modelo seleccionado con el fin de determinar cuántos y cuáles parámetros pueden ser recuperados del modelo.

Luego se presentan los resultados de la estimación de los parámetros con el modelo seleccionado y las condiciones númericas adicionales impuestas, tales como el término de regularización, el ajuste del volumen a las medidas de Densidad Óptica y el ajuste del modelo a los datos correspondientes a las mediciones en dos de las concentraciones.

Y finalmente, se presentan los árboles filogénicos conseguidos con los diferentes métodos.

3.1. Selección de Modelos

Para hacer uso de los criterios AIC_c y BIC y hacer la selección del modelo que mejor describe los datos se hicierón experimentos numéricos con las cepas 1 a 5 en la concentración 4 (7.8 μM). Los resultados se presentan a continuación. (El algoritmo se corrió con 500 puntos y 2000 iteraciones)

Los resultados que se presentan en las tablas corresponden al valor de la función de verosimilitud ($-2ln(\widehat{L}_i)$) y los resultados de los criterios (AIC_c) y (BIC) explicados anteriormente. En cada tabla se resumen los valores obtenidos para una misma cepa con los diferentes modelos.

Cepa 1

Modelo	Cepa	$-2ln(\widehat{L}_i)$	AIC_c	BIC
1	1	$9{,}837000e+03$	$9{,}8471052e+03$	$9{,}837000e+03$
2	1	$4{,}352000e+03$	$4{,}414400e+03$	$4{,}310000e+03$
3	1	$4{,}046000e+03$	$4{,}090000e+03$	$4{,}080000e+03$
4	1	$4{,}076000e+03$	$4{,}138400e+03$	$4{,}069000e+03$
5	1	$3{,}670671e+03$	$3{,}707594e+03$	$3{,}702451e+03$

Cepa 2

Modelo	Cepa	$-2ln(\widehat{L}_i)$	AIC_c	BIC
1	2	$1{,}060496e+04$	$1{,}061506e+04$	$1{,}061767e+04$
2	2	$1{,}031602e+04$	$1{,}037842e+04$	$1{,}027470e+04$
3	2	$1{,}060496e+04$	$1{,}064896e+04$	$1{,}057000e+04$
4	2	$1{,}060496e+04$	$1{,}122896e+04$	$1{,}059860e+04$
2	2	$7{,}520984e+03$	$7{,}557907e+03$	$7{,}552764e+03$

Cepa 3

Modelo	Cepa	$-2ln(\widehat{L}_i)$	AIC_c	BIC
1	3	$6{,}423286e+03$	$6{,}433912e+03$	$6{,}435998e+03$
2	3	$7{,}314108e+03$	$7{,}376508e+03$	$7{,}272793e+03$
3	3	$6{,}423286e+03$	$6{,}473286e+03$	$6{,}388328e+03$
4	3	$1{,}739882e+04$	$1{,}736122e+04$	$1{,}739246e+04$
5	3	$4{,}574864e+03$	$4{,}611787e+03$	$4{,}606645e+03$

Cepa 4

Modelo	Cepa	$-2ln(\widehat{L}_i)$	AIC_c	BIC
1	4	$4{,}396695e+04$	$4{,}397705e+04$	$4{,}397966e+04$
2	4	$4{,}559966e+04$	$4{,}566206e+04$	$4{,}555834e+04$
3	4	$4{,}396695e+04$	$4{,}401095e+04$	$4{,}393199e+04$
4	4	$4{,}551575e+04$	$4{,}557815e+04$	$4{,}548079e+04$
5	4	$4{,}318859e+04$	$4{,}355782e+04$	$4{,}322037e+04$

Cepa 5

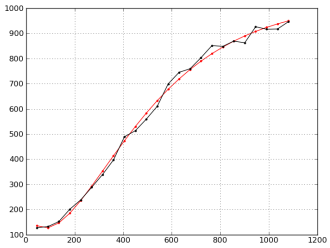
Cepa	$-2ln(\widehat{L}_i)$	AIC_c	BIC	
1	5	$7{,}677076e+03$	$7{,}687181e+03$	$7{,}689782e+03$
2	5	$1{,}120787e+04$	$1{,}127027e+04$	$1{,}116656e+04$
3	5	$7{,}677076e+03$	$7{,}721076e+03$	$7{,}642117e+03$
4	5	$2{,}535636e+04$	$2{,}541876e+04$	$2{,}532140e+04$
5	5	$5{,}170459e+03$	$4{,}970459e+03$	$5{,}2022404e+03$

Si comparamos los valores de AIC_c y BIC de cada una de las tablas, para cada una de las cepas se puede ver que, en la mayoría de los casos, los mejores resultados se consiguen para el modelo cinco. Dado que el criterio de decisión dice que es mejor modelo aquel cuyo valor de AIC_c y BIC sea menor, concluimos que el Modelo de Musky es el que mejor describe los datos.

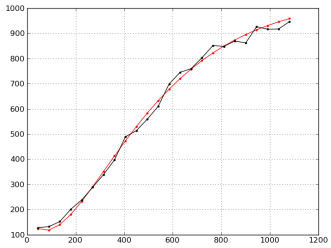
3.1.1. Ajuste de los Modelos a los Datos Experimentales

A continuación se presentan los ajustes de los modelos para la cepa 3 en la concentración 4, el eje x corresponde a los niveles de producción de proteína (GFP) y el eje y corresponde al tiempo (t). Puede verse gráficamente que el modelo cinco se ajusta mejor a los datos.

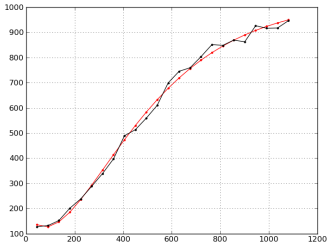
Modelo 1



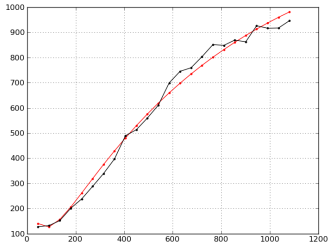
Modelo 2



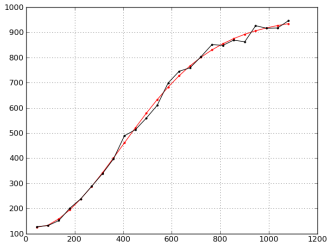
Modelo 3



Modelo 4



Modelo 5

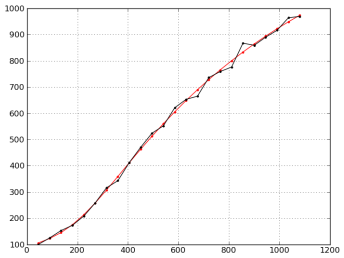


Cuadro 3.1: Ajustes de los Modelos a los Datos Experimentales

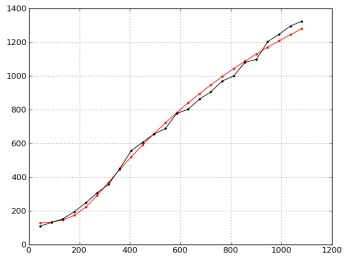
Dado que éste es el que ofrece una mejor descripción de los datos, será el que se usará para el problema de estimación de parámetros.

Antes de esto, se hizo un ejercicio adicional, se tomarón los parámetros obtenidos para las cepas 1 y 3 en la concentración 4 y se fijarón, con la intensión de ver si era posible predecir resultados en otras concentraciones. Consiguiendo un resultado positivo como lo muestran las siguientes gráficas.

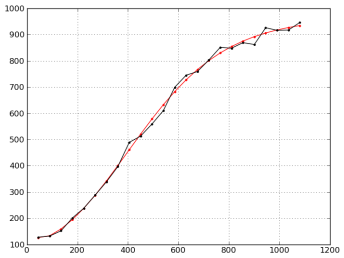
Cepa 1 concentracón 4



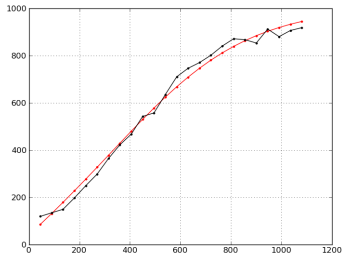
Predicción para concentración 5



Cepa 3 concentración 4



Predicción para concentración 3



Cuadro 3.2: Predicción para otras concentraciones con un conjunto de parámetros fijos

3.2. Análisis cualitativo

Los modelos considerados tienen un único estado estacionario, el cual es un atractor si la concentración de inductor es positiva. Si el nivel de inductor es cero, se sigue presentando un estado estacionario asintóticamente estable que da cuenta de una producción basal del represor y la proteína.

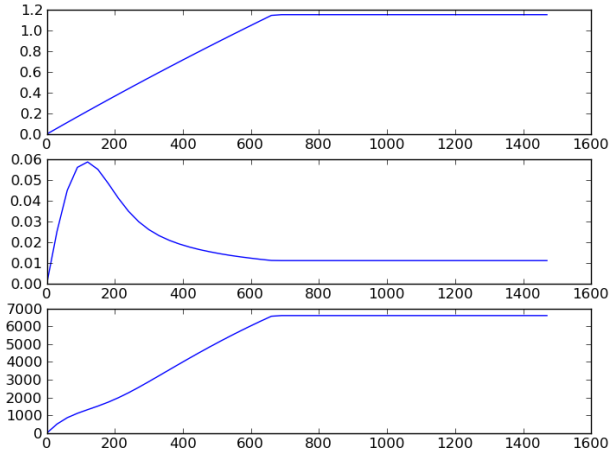


Figura 3.1: Solución del Modelo 5

La primera y tercer gráfica corresponden al Inductor y a la producción de proteína, estos alcanzan un punto de saturación como se observa en la gráfica y como veremos mas adelante. Es un resultado coherente con la realidad, dado que, por más inductor que se tenga en el medio, la célula solo tiene capacidad para una cantidad limitada del mismo.

La segunda gráfica corresponde al represor, su forma se debe a que, no sólo se tiene pérdida por degradación natural, sino que, también hay pérdida cuando los represores se unen con el inductor, por tanto, mientras mas IPTG intracelular

se tenga, mayor será la pérdida de represores.

1. **Puntos Críticos en Presencia de Inductor**

$$r(I_{max} - I) = 0$$
$$k_l - (\delta^0 + \delta^1 I)R = 0$$
$$\frac{k_p}{1 + \alpha R^2} - \gamma_p P = 0$$

El único punto estacionario en presencia de inductor es:

$$I = I_{max}, \quad R = \frac{k_l}{\delta^0 + \delta^1 I_{max}}, \quad P = \frac{k_p(\delta^0 + \delta^1 I_{max})^2}{\gamma_p((\delta^0 + \delta^1 I_{max})^2) + \alpha {k_l}^2}$$

Al calcular la matriz Jacobiana y evaluarla en el punto se obtiene:

$$J(I, R, P) = \begin{pmatrix} -r & 0 & 0 \\ \frac{-\delta^1 k_l}{\delta^0 + \delta^1 I_{max}} & -(\delta^0 + \delta^1 I_{max}) & 0 \\ 0 & \frac{-2\alpha k_p k_l (\delta^0 + \delta^1 I_{max})}{((\delta^0 + \delta^1 I_{max})^2 + \alpha k_l^2)} & -\gamma_p \end{pmatrix}$$

Eigenvalores:

$$\lambda_1 = -r,\ \lambda_2 = -(\delta^0 + \delta^1 I_{max}),\ \lambda_3 = -\gamma_p$$

Dado que las constantes cinéticas son positivas se sigue que el punto (I,R,P) es asintóticamente estable. Los valores negativos para las constantes carecen de significado físico.

2. **Puntos Críticos en Ausencia de Inductor**

Para $I_{max} = 0$ el punto estacionario está dado por:

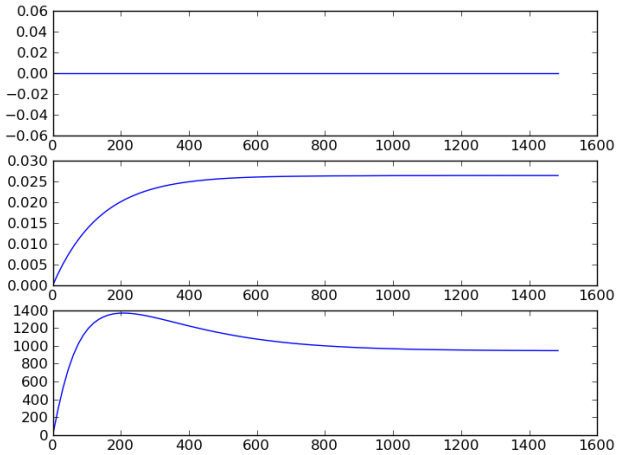


Figura 3.2: Solución del Modelo sin IPTG

$$I = 0,\ R = \frac{k_l}{\delta^0},\ P = \frac{k_p(\delta^0)^2}{\gamma_p((\delta^0)+\alpha {k_l}^2}$$

Éste también es asintóticamente estable.

Aunque el operón presenta su mayor actividad en presencia de un inductor, tiene una producción basal en ausencia del mismo. Este resultado es el que estamos observando cuando tomamos $I_{max} = 0$, la forma de la solución de la proteína se debe a que, cuando aumenta la cantidad de represores crece la probabilidad de que el operador esté ocupado, disminuyendo la producción de GFP.

3.3. Análisis de Sensibilidad

El análisis de sensibilidad del modelo se llevo a cabo numéricamente. En un entorno de los valores numéricos de los parámetros hallados mediante la solución del problema inverso haciendo uso de la librería Sundials- Python.

r	Ritmo al que ingresa el IPTG al interior de la célula
k_l	Máximo ritmo de iniciación de la transcripción de lacI
δ^0	Degradación de lacI
δ^1	Asociación de lacI con el IPTG
k_p	Ritmo de traducción de la proteína
α	Fuerza de ocupación de lacI
γ_p	Degradación de la proteína
I_0	Condición inicial para IPTG
R_0	Condición inicial para represores
P_0	Condición inial para la proteína

Ecuaciones de Sensibilidad para el Modelo 5

Parámetros $p = (r, k_l, \delta^0, \delta^1, k_p, \alpha, \gamma_p, I_0, R_0, P_0)$

$$\dot{s_i} = J.s_i = \begin{pmatrix} -r & 0 & 0 \\ -\delta^1 R & -(\delta^0 + \delta^1 I) & 0 \\ 0 & \frac{-2\alpha k_p R}{(1+\alpha R^2)^2} & -\gamma_p \end{pmatrix} \frac{dy}{dp_i}$$

Derivadas del sistema con respecto a los parámetros.

$\frac{df}{dr} = \begin{pmatrix} I_{max} - I \\ 0 \\ 0 \end{pmatrix}$, $\frac{df}{dk_l} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\frac{df}{d\delta^0} = \begin{pmatrix} 0 \\ -R \\ 0 \end{pmatrix}$, $\frac{df}{d\delta^1} = \begin{pmatrix} 0 \\ -RI \\ 0 \end{pmatrix}$,

$\frac{df}{dk_p} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{(1+\alpha R^2)} \end{pmatrix}$, $\frac{df}{d\alpha} = \begin{pmatrix} 0 \\ 0 \\ \frac{-k_p R^2}{(1+\alpha R^2)^2} \end{pmatrix}$, $\frac{df}{d\gamma_p} = \begin{pmatrix} 0 \\ 0 \\ -P \end{pmatrix}$

3.3.1. Algunos Resultados numéricos

Dado el conjunto de parámetros para la cepa 1 en la concentración 4:

$$[r : 1{,}862e^{-03}, k_l : 6{,}1e^{-05}, \delta^0 : 5e^{-06}, \delta^1 : 2{,}549e^{-03},$$

$$k_p : 1{,}3670490e^{01}, \alpha : 1{,}23e^{-04}, \gamma_p : 1{,}134e^{-03}, I_0 : 2{,}93473e^{-01},$$

$$R_0 : 8{,}5449331e^{01}, P_0 : 1{,}32366001e^{02}]$$

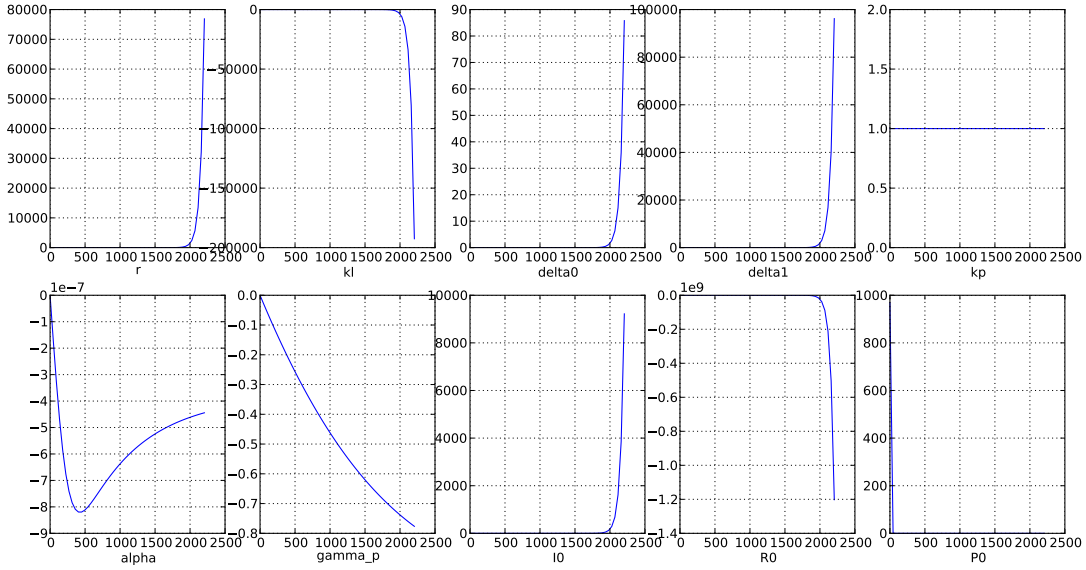


Figura 3.3: Resultados de Sensibilidad cepa 1

Considerando un nuevo conjunto de parámetros, ahora para la cepa 2 en la concentración 4:

$$[r : 9{,}014e^{-03}, k_l : 1e^{-07}, \delta^0 : 7{,}8e^{-05}, \delta^1 : 6{,}36e^{-04},$$
$$k_p : 1{,}9291e^{01}, \alpha : 1{,}965703e^{02}, \gamma_p : 3{,}198e^{-03}, I_0 : 2{,}18502e^{-01},$$
$$R_0 : 8{,}9099e^{01}, P_0 : 1{,}49237e^{-01}]$$

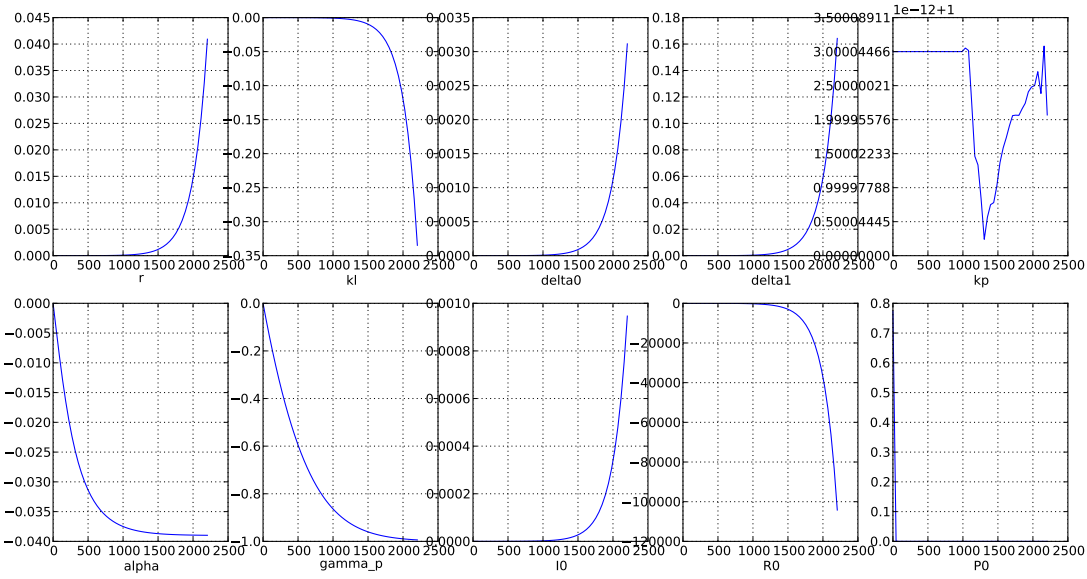


Figura 3.4: Resultados de Sensibilidad cepa 2

(Ver otros resultados en el Anexo)

De estos resultados se observa que el modelo es sensible con respecto a r, δ^0, δ^1 y I_0, los datos graficados corresponden a $\frac{dP}{dp_i}$, así que es coherente lo que se obtiene, dado que cambios en estos parámetros hacen que la producción de

la proteína se vea afectada. Estos parámetros están relacionados directamente con el estado de los represores, quienes juegan un papel crucial en la inducción del operón.

Aunque el conjunto de parámetros sea cambiado, la sensibilidad del modelo se mantiene con respecto a los mismas constantes.

El objetivo de este análisis es determinar cuáles parámetros son más influyentes en el modelo, dado que pequeños cambios en estos pueden producir cambios significativos en la solución del sistema. Teniendo esta información se pueden diseñar experimentos en los que se tenga un mejor control o manejo de dichas constantes.

Por otra parte, al realizar el análisis del mapeo genotipo-fenotipo, se puede analizar si las variaciones fenotípicas están relacionadas directamente con estas constantes o no. Además, conocer las matrices de sensibilidad nos permitirán hacer estudios posteriores sobre la identificabilidad de los parámetros.

3.4. Identificabilidad de Parámetros

En esta sección se presentan los resultados obtenidos para la Identificabilidad de los parámetros, en la primer parte se muestran los resultados análiticos y posteriormente algunos resultados numéricos con base en las matrices de sensibilidad de arriba.

Dados los vectores de parámetros

$$\hat{p} = [\hat{r}, \hat{k}_l, \hat{\delta}^0, \hat{\delta}^1, \hat{k}_p, \hat{\alpha}, \hat{\gamma}_p, \hat{I}_0, \hat{R}_0, \hat{P}_0]$$
$$p = [r, k_l, \delta^0, \delta^1, k_p, \alpha, \gamma_p, I_0, R_0, P_0]$$

Con el fin de aplicar el teorema de la expansión en series de Taylor, calculamos las siguientes derivadas

$$\begin{aligned}
a_0(p) &= P_0 \\
a_1(p) &= P' = \frac{k_p}{1+\alpha R^2} - \gamma_p P_0 \\
a_2(p) &= (\alpha^2 {\gamma_p}^2 P {R_0}^4 + 2\alpha {\gamma_p}^2 P {R_0}^2 + 2\alpha k_p \delta^1 I {R_0}^2 + 2\alpha k_p \delta^0 {R_0}^2 \\
&\quad - \alpha \gamma_p k_p {R_0}^2 - 2\alpha k_l k_p R_0 + {\gamma_p}^2 P_0 - \gamma_p k_p)/\left(\alpha {R_0}^2 + 1\right)^2
\end{aligned}$$

$$
\begin{aligned}
a_3(p) = {} & 2\alpha^3{\gamma_p}^3 P {R_0}^6 - 6\alpha^2{\gamma_p}^3 P_0 {R_0}^4 + 4\alpha^2 k_p r \delta^1 I_{max} {R_0}^4 \\
& + 8\alpha^2 k_p {\delta^1}^2 {I_0}^2 {R_0}^4 + 16\alpha^2 k_p \delta^0 \delta^1 I {R_0}^4 - 4\alpha^2 k_p r \delta^1 I_0 {R_0}^4 \\
& - 4\alpha^2 \gamma_p k_p \delta^1 I R^4 + 8\alpha^2 k_p {\delta^0}^2 R^4 - 4\alpha^2 \gamma_p k_p \delta^0 R^4 \\
& + 2\alpha^2 {\gamma_p}^2 k_p {R_0}^4 - 20\alpha^2 k_l k_p \delta^1 I {R_0}^3 - 20\alpha^2 k_l k_p \delta^0 {R_0}^3 \\
& + 4\alpha^2 \gamma_p k_l k_p {R_0}^3 - 6\alpha {\gamma_p}^3 P {R_0}^2 + 4\alpha k_p r \delta^1 I_{max} {R_0}^2 - 8\alpha k_p {\delta^1}^2 {I_0}^2 {R_0}^2 \\
& - 16\alpha k_p \delta^0 \delta^1 I_0 {R_0}^2 - 4\alpha k_p r \delta^1 I_0 {R_0}^2 - 4\alpha \gamma_p k_p \delta^1 I_0 {R_0}^2 - 8\alpha k_p {\delta^0}^2 {R_0}^2 \\
& - 4\alpha \gamma_p k_p \delta^0 {R_0}^2 + 12\alpha^2 {k_l}^2 k_p {R_0}^2 + 4\alpha {\gamma_p}^2 k_p {R_0}^2 + 12\alpha k_l k_p \delta^1 I_0 R_0 \\
& + 12\alpha k_l k_p \delta^0 R_0 + 4\alpha \gamma_p k_l k_p R_0 - 2{\gamma_p}^3 P_0 \\
& - 4\alpha {k_l}^2 k_p + 2{\gamma_p}^2 k_p / (\alpha^3 {R_0}^6 + 3\alpha^2 {R_0}^4 + 3\alpha {R_0}^2 + 1)
\end{aligned}
$$

Comparando término a término se consigue que

$$P_0, \gamma_p, k_p, \delta^0, r, \delta^1, I_0$$

son identificables al igual que los productos $\alpha {R_0}^2$ y $\alpha k_l R_0$ y $\alpha {k_l}^2$

Es decir que el sistema no es completamente identificable si se tiene solo la medida de la proteína, éste será identificable sólo si se tienen condiciones adicionales.

Un resultado similar fue obtenido por Munsky (Ver información suplementaria de Munsky 2009). Sin embargo, esto puede ser superado dado el conjunto de datos experimentales y algunas herramientas computacionales [5] o fijando uno de los parámetros que aparece en los productos.

La identificabilidad de los parámetros no depende sólo del modelo propuesto, sino que también depende fuertemente de los datos experimentales que se tienen, dado que la estimación de estos se hace por ajustes de la estructura del modelo a datos experimentales, por lo que, un modelo estructuralmente identificable puede no exhibir identificabilidad práctica debido al ruido de los datos o a la cantidad que se tiene de los mismos. [52]

Para una identificabilidad total de los parámetros son necesarias $2p+1$ mediciones en el tiempo [5, 25], suponiendo que los datos están libres de ruido. En este caso, se quiere conseguir la identificación de 10 parámetros y se cuenta con un total de 24 medidas en el tiempo en 12 concentraciones de IPTG diferentes, es decir que se cuenta con una buena colección de datos para la identificación.

El análisis de identificabilidad práctica obedece a la solución del problema inverso que se propone mas adelante.

3.5. Solución del Problema Inverso

Aunque analtíamente vemos que no es posible recuperar todos los parámetros, se pueden dar algunas condiciones adicionales a la hora de resolver el problema inverso, que permitan estabilizar los valores de los parámetros y obtener resultados confiables.

A continuación se presentan algunas de las estrategias exploradas y algunos resultados numéricos. En todos los casos se tuvierón las siguientes consideraciones:

- Se fijó el valor de la degradación de la proteína.
- Se usó un término de regularización para resolver el problema de optimización.
- Se usó un algoritmo de Evolución.
- Se acotó la búsqueda de los parámetros a intervalos teniendo en cuenta los valores encontrados en la literatura. (Ver valores de la literatura en el apéndice)
- El algoritmo se corrió con un total de 500 puntos y 5000 iteraciones.

3.5.1. Estrategia 1 : Una Concentración y Término de Regularización

La propuesta es utilizar los datos de GFP para una sola concentración de IPTG para la solución del problema inverso e imponer un término de regularización.

La concentración utilizada es la concentración cuatro, dado que, es una concentración que se aproxima más a las concentraciones en las que la bacteria se puede encontrar en un ambiente silvestre. Además, se resolvió el problema con las concentraciones 3, 4, 5, 6 y 9 y el mejor resultado se obtuvo siempre con la concentración 4, en otras concentraciones los niveles se tiene mayor ruido en las lecturas de la fluorescencia.

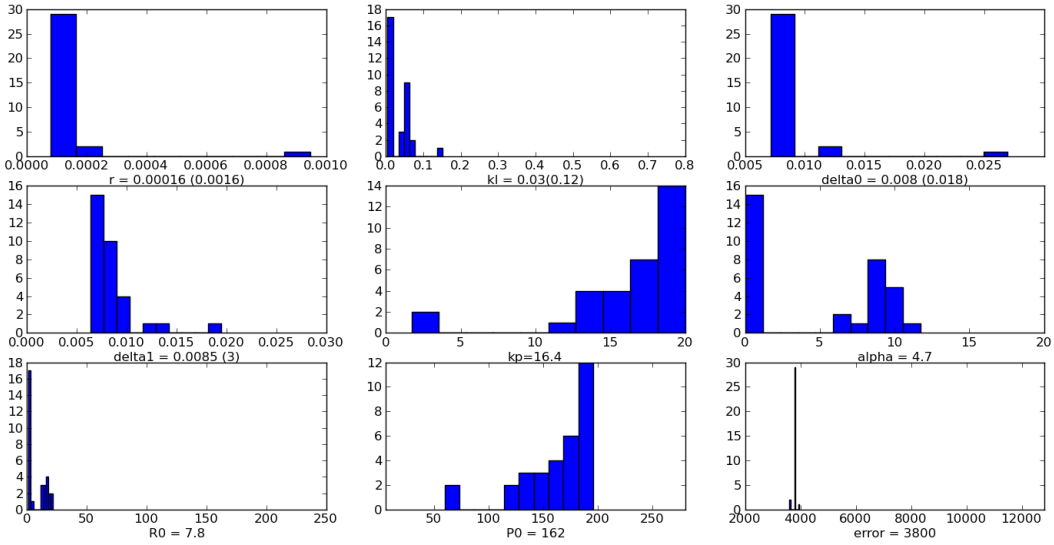
El término a optimizar es:

$$\min_{p} ||F(p) * V(p) - GFP||_Y^2 + \alpha R(p, p_0)$$

Algunos resultados numéricos:

Cepa	Sin regularización	con regularización
1	$3{,}670671e+03$	$3{,}656889e^{+}03$
2	$7{,}520984e+03$	$7{,}512301e+03$
3	$4{,}574865e+03$	$4{,}573017e+03$
4	$4{,}318859e+04$	$4{,}283054e+03$
5	$5{,}170460e+03$	$5{,}151809e+03$

De los resultados se puede ver que el ajuste mejora (mejor score) cuando se implementa el término de regularización. Para la cepa 1 se corrió el código 32 veces con 500 puntos y 5000 interacciones para observar la distribución de los parámetros.



Aunque se halla impuesto el término de regularización, algunos parámetros presentan variación en intervalos muy amplios, es el caso de k_p, α, P_0 y R_0.

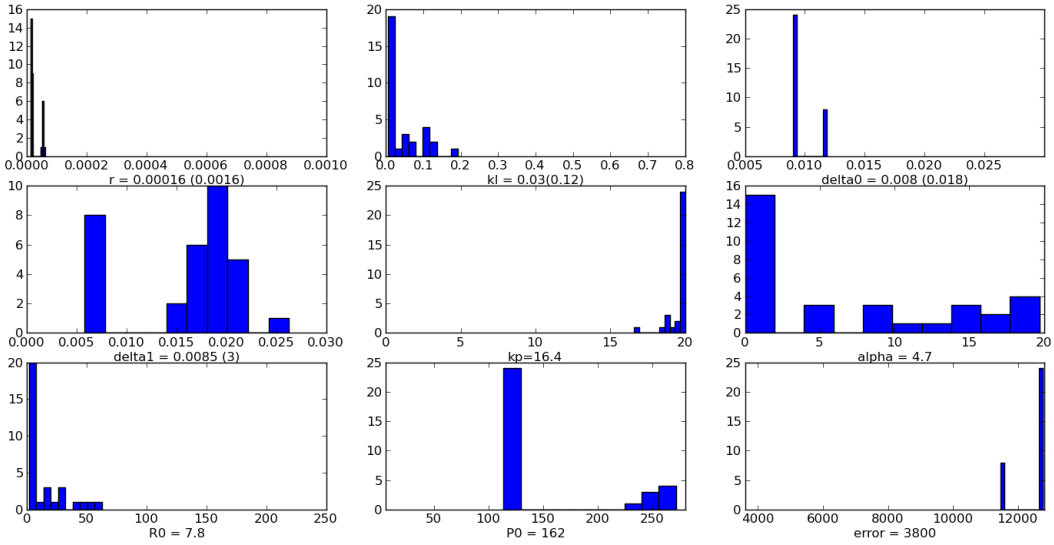
3.5.2. Estrategia 2 : Dos Concentraciones

Dado que se tienen datos para el mismo fenómeno en diferentes concentraciones se propone usar dos de estas para realizar el ajuste, esto con el objetivo de minimizar la desviación estándar.

El término a minimizar es:

$$\min_p ||F(p) * V(p) - Datos_1||_Y^2 + \min_p ||F(p) * V(p) - Datos_2||_Y^2 + \alpha R(p, p_0)$$

Cepa	Sin Regularización	Estrategia 1	Estrategia 2
1	$3{,}670671e+03$	$3{,}656889e^{+}03$	$1{,}340946e+04$
2	$7{,}520984e+03$	$7{,}512301e+03$	$3{,}225273e+04$
3	$4{,}574865e+03$	$4{,}573017e+03$	$1{,}076294e+04$
5	$5{,}170460e+03$	$5{,}151809e+03$	$1{,}477248e+04$



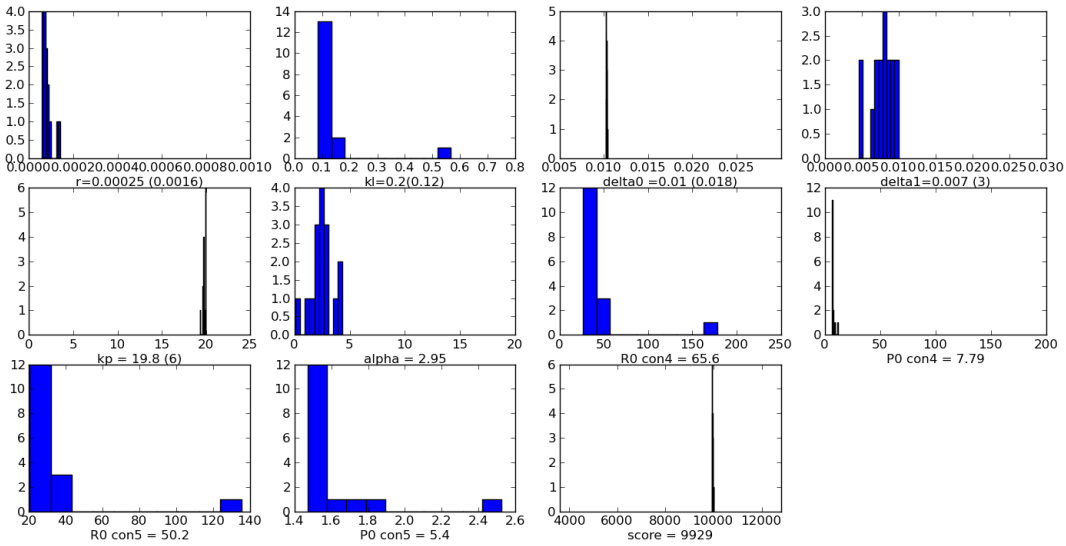
3.5.3. Estrategia 3: Dos Concentraciones y Ajuste de los Parámetros con OD

Además de usar los datos correspondientes a dos concentraciones distintas, se usarán los datos de OD, dado que se sabe que existe una relación entre la densidad óptica y la cantidad de células en un tiempo t ($OD \approx \beta V$), usando este hecho se ajustan los parámetros del volumen celular con los datos de densidad óptica.

Nuevamente se corre 32 veces el código para la cepa uno, las concentraciones

Cepa	Sin regularización	Estrategia 1	Estrategia 2	Estrategia 4
1	$3{,}670671e+03$	$3{,}656889e^{+}03$	$3{,}820059e+03$	$1{,}17286e+04$
2	$7{,}520984e+03$	$7{,}512301e+03$	$7{,}418519e+03$	$3{,}12255e+04$
3	$4{,}574865e+03$	$4{,}573017e+03$	$5{,}420524e+03$	$1{,}12700e+04$
4	$4{,}318859e+04$	$4{,}283054e+03$	$2{,}696835e+04$	$1{,}40095e+05$

que se están usando para los ajustes son la 4 y la 5.

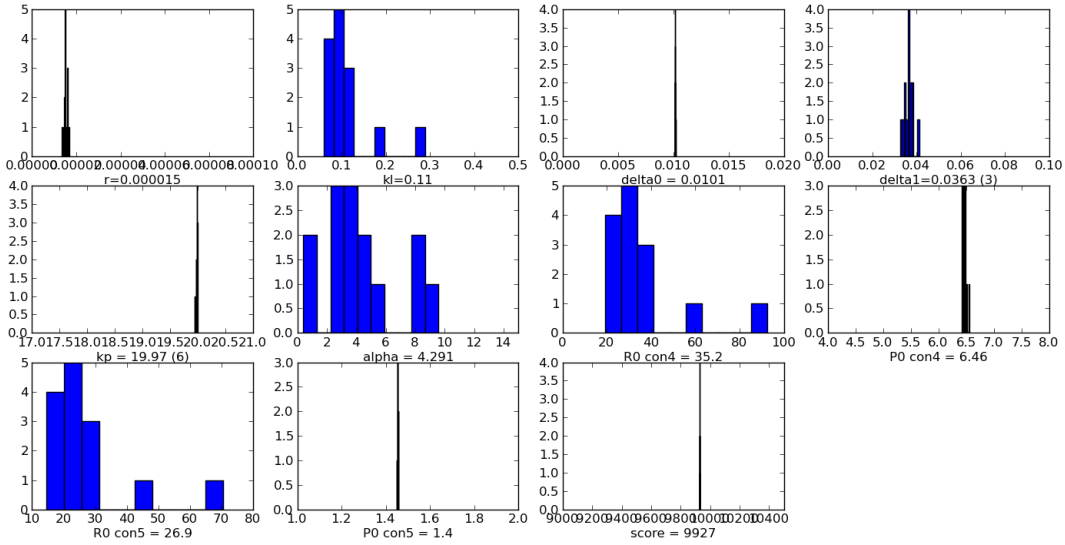


Al comparar los resultados, se puede ver que con la estrategia dos se disminuye la varianza en algunas de las constantes, comparada con la estrategia uno, es el caso de δ^0, k_p, y P_0. Sin embargo, con la tres se mejoran las desviaciones considerablemente comparada con las otras dos.

El aumento en el error se debe a que se obliga al modelo a ajustarse a muchos mas datos, pero aunque se pierde un poco en éste, se gana estabilidad en

los parámetros y es finalmente lo que se quiere conseguir.

Los resultados anteriores se consiguen corriendo el código con 10000 iteraciones sin condiciones extras para detener la DE, pero cuando inponemos una condición para detenerla los resultados que se obtienen son los siguientes.



Aquellos parámetros que aún muestran una variación corresponden a los parámetros que no son identificables, es necesario fijar uno de ellos para conseguir la identificabilidad de los otros dos, en este caso, podría ser conveniente fijar el valor de R_0.

Es de notar que los parámetros que se obtienen al resolver el problema inverso no tienen que ser iguales a los reportados en la literatura, dado que,

corresponden a cepas de bacterias diferentes, sin embargo se espera que sean valores aproximados.

3.5.4. Árboles Filogenéticos

El primer paso para conseguir el árbol para Neighbor Joining es determinar cuál es la mejor distancia para comparar las secuencias, dentro de estas se puede encontrar Jukes-Cantor, Kimura 2, Felstein, entre otras, en Mega se tiene la opción, una vez se tienen las secuencias, de correr un código que determinará a partir de los criterios de AIC y BIC cuál es la distancia más apropiada para realizar el análisis filogenético.

Al hacer esto, se consiguen los siguientes resultados:

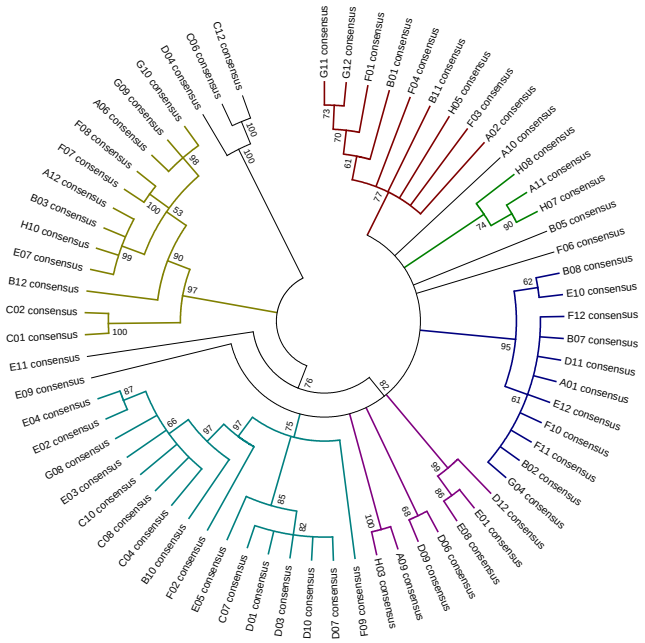
Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)
K2+G+I	134	6783.810	5602.914	-2667.094	0.05	0.39	2.11	0.250	0.250	0.250	0.250
K2+G	133	6787.776	5615.688	-2674.486	n/a	0.97	1.96	0.250	0.250	0.250	0.250
K2+I	133	6803.344	5631.256	-2682.270	0.05	n/a	2.09	0.250	0.250	0.250	0.250
JC+G+I	133	6814.426	5642.338	-2687.811	0.00	0.33	0.50	0.250	0.250	0.250	0.250
K2	132	6821.338	5658.057	-2696.676	n/a	n/a	2.08	0.250	0.250	0.250	0.250
JC+G	132	6823.673	5660.392	-2697.843	n/a	2.13	0.50	0.250	0.250	0.250	0.250
JC	131	6825.703	5671.229	-2704.267	n/a	n/a	0.50	0.250	0.250	0.250	0.250
T92+G+I	135	6833.848	5644.145	-2686.704	0.06	0.40	2.13	0.229	0.229	0.271	0.271
JC+I	132	6834.305	5671.024	-2703.160	0.05	n/a	0.50	0.250	0.250	0.250	0.250
T92+G	134	6838.955	5658.060	-2694.666	n/a	0.92	1.86	0.229	0.229	0.271	0.271
T92	133	6850.663	5678.575	-2705.929	n/a	n/a	2.08	0.229	0.229	0.271	0.271
T92+I	134	6853.451	5672.555	-2701.914	0.06	n/a	2.09	0.229	0.229	0.271	0.271
HKY+G+I	137	6856.635	5649.318	-2687.279	0.06	0.40	2.12	0.226	0.233	0.270	0.271
HKY+G	136	6861.849	5663.339	-2695.295	n/a	0.92	1.86	0.226	0.233	0.270	0.271
TN93+G+I	138	6864.874	5648.750	-2685.990	0.06	0.38	2.12	0.226	0.233	0.270	0.271
HKY	135	6873.502	5683.800	-2706.531	n/a	n/a	2.08	0.226	0.233	0.270	0.271
HKY+I	136	6876.217	5677.707	-2702.479	0.06	n/a	2.09	0.226	0.233	0.270	0.271
GTR+G	140	6877.385	5643.647	-2681.427	n/a	0.40	1.92	0.226	0.233	0.270	0.271
TN93	136	6881.278	5682.768	-2705.010	n/a	n/a	2.06	0.226	0.233	0.270	0.271
TN93+I	137	6884.289	5676.972	-2701.106	0.06	n/a	2.09	0.226	0.233	0.270	0.271

Lo que quiere decir es que la distancia de Kimura 2 es la más apropiada dado el criterio BIC. La letra G indica una distribución gamma y la I es la

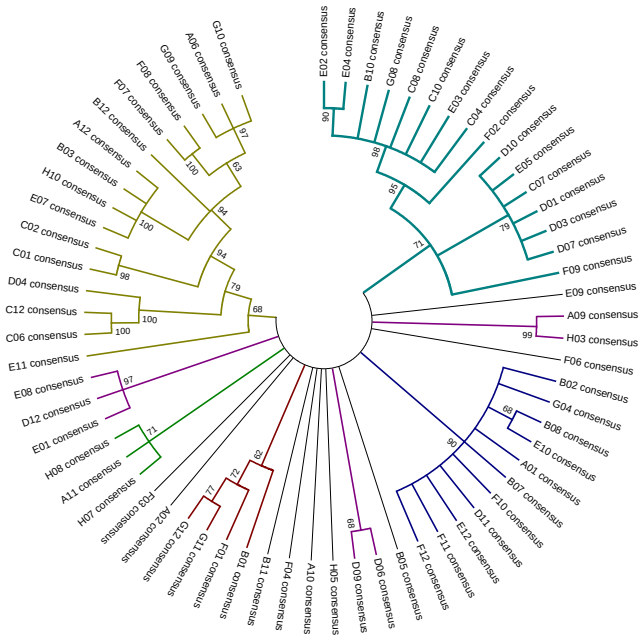
proporción de sitios invariantes (Tasa de evolución = 0), esta información es usada a la hora de construir el árbol.

3.5.5. Neighbor Joining

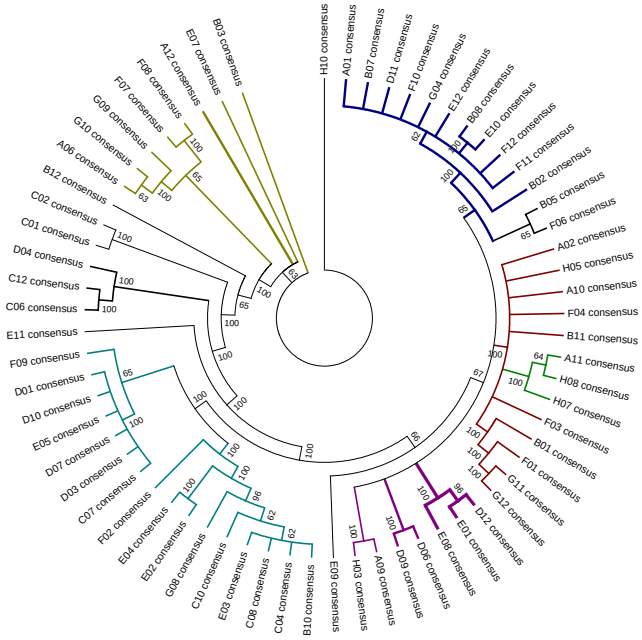


Para los casos de Máxima Verosimilitud y Parsimonia no es necesario hacer la elección de una medida de distancia dado que para estos árboles se emplean directamente las secuencias.

3.5.6. Máxima Verosimilitud



3.5.7. Parsimonia



Al comparar los diferentes árboles vemos que coinciden en la mayoría de los clados o ramas, esto indica que la filogenia es consistente. Los clados iguales aparecen coloreados del mismo tono. Aquellas secuencias que se encuentran en una misma rama pertenecen al mismo grupo, lo que queda es comparar si coincide o no con los agrupamientos de los diferentes parámetros.

Con este análisis se consigue un primer agrupamiento de las secuencias con las que se puede trabajar, los grupos obtenidos son:

- **Grupo 1**: C01, C02, B12, E07, H10, B03, A12, F07, F08, A06, G09, G10
- **Grupo 2**: C12, C06, D04
- **Grupo 3**: G11, G12, F01, B01, F04, B11, H05, F03, A02
- **Grupo 4**: H08, A11, H07
- **Grupo 5**: B08, E10, F12, B07, D11, A01, E12, F10, B02, G04
- **Grupo 6**: D12, E01, E08, D06, D09, A09, H03
- **Grupo 7**: F09, D07, D10, D03, D01, C07, E05, F02, B10, C04, C08, C10, E03, G08, E02, E04

Las cepas E11, E09, B05, y F06 no se encuentran en ninguno de estos grupos.

Capítulo 4

Discusión y Conclusiones

Se analizarón cinco modelos que describen la dinámica del operón con diferente nivel de detalle, como se ha mencionado durante todo el trabajo, uno de los intereses es elegir aquel que mejor describa los datos, pero existen otros motivos que cobran importancia a la hora de elegir el modelo.

Como se puede ver en los resultados, al aplicar los criterios de teoría de la información se obtiene que el modelo propuesto por Munsky acoplado con un crecimiento logístico para las bacterias es el que mejor se ajusta a los datos, sin embargo, esto no es suficiente para resolver el problema.

Se quiere un modelo que involucre la mayor cantidad posible de constantes cinéticas relacionadas con las regiones del operón. Analizando cada uno de los modelos se tiene:

Modelo 1: Es uno de los modelos más simples, su simplicidad permite realizar análisis teóricos con mayor facilidad, pero éste solo ofrece información

de cuatro constantes, de las cuales, sólo una es de real interés, k_M (tasa de transcripción), esta constante se encuentra directamente relacionada con el promotor del operón, en tanto que γ_M, γ_p y k_p que corresponden a la degradación del mRNA, degradación de GFP y tasa de transcripción respectivamente, no se encuentran relacionadas directamente con procesos del operón; así que, podríamos decir que es un poco pobre la información que se puede recuperar de este modelo.

Modelo 2, 3 y 4: Estos modelos involucran un poco más de detalle de la dinámica, k_M: Tasa de transcripción, relacionada con la región promotora, K_I: Afinidad represor - inductor, relacionada con la región de lac I, k_R: Afinidad represor - operador, involucrada con la región de lac I y con la zona del operador. Se tienen constantes asociadas a cada uno de las regiones de interés, el punto es, cuántas son identificables o se pueden recuperar de manera confiable numéricamente.

Cuando se hacen los cálculos para la aplicación de teoremas de identificabilidad es casi imposible recuperar los parámetros, los cálculos se hacen imposibles, ahora, si se intentan recuperar numéricamente no es posible estabilizar el valor de K_I, así que, aunque estos modelos incluyen una cantidad considerable de constantes, su análisis analítico no es viable y numéricamente no se obtienen resultados mejores que los que se consiguen con otros modelos.

Modelo 5: Finalmente se tiene el modelo de Munsky, el cual, logra explicar mejor los datos que los demás modelos y sus constantes permiten recuperar información de las regiones de interés en el operón. Se tiene: δ^0, δ^1 y k_l asociadas

con la region lacI y α asociada a la región del operador, con éste modelo se pierde información de la región promotora pero se tiene información confiable de las otras. Por otra parte, al realizar el análisis de identificabilidad se consigue que son identificables todas las constantes excepto tres, situación que se puede controlar imponiendo condiciones adicionales, por otra parte, el análisis cualitativo y de sensibilidad del modelo deja ver que éste está dando cuenta de la dinámica del operón.

Con el análisis cualitativo se obtiene un punto estacionario, tanto en presencia como en ausencia de inductor; en ambos casos dicho punto da cuenta de condiciones naturales, en el primero muestra un punto de saturación, lo que es normal, dado que, aunque la célula esté expuesta a grandes cantidades de alimento tiene un límite de consumo y en ausencia de inductor se da una producción basal de proteína y es lo que indica el punto estacionario en este caso.

En cuanto al análisis de sensibilidad, como se mencionó anteriormente, el modelo es sensible a r, tasa a la que ingresa el IPTG al interior de la célula, δ^0 y δ^1 que tienen que ver con la pérdida del represor, bien sea por degradación natural o por represión y finalmente, es sensible a las condiciones iniciales de IPTG (I_0), lo cual es coherente dada la dinámica del operón y la producción de la proteína, así que, tenemos un modelo que da cuenta del fenómeno.

Lo anterior implica que el modelo 5

- Explica la dinámica del operón dado el análisis de sensibilidad y el cualitativo.

- Se ajusta bien a los datos.
- Permite recuperar una cantidad admisible de constantes que ofrecen información de la cinética del operón.

Una vez identificado el modelo, es de interés encontrar una estrategia numérica que permita resolver de manera confiable el problema inverso. Se estudiarón las siguientes opciones:

- La primera fué usar un término de regularización teniendo en cuenta la propuesta de Engl [20].
- Se ajustó el modelo con dos concentraciones diferentes de IPTG y usando el término de regularización.
- Finalmente se ajustó el modelo a dos concentraciones de IPTG y a los datos de OD usando el término de regularización.

De los resultados obtenidos para la cepa 1, se puede concluir que realizar el ajuste usando datos de dos de las concentraciones, específicamente la cuatro y la cinco, y las medidas de OD es una buena estrategia, dado que, estabiliza mejor los parámetros, es decir, la desviación estándar mejora con esto, es menor que en los casos anteriores, dando una mayor confiabilidad en los datos resultantes.

Algo por intentar es, en lugar de tomar datos de GFP en diferentes concentraciones, realizar experimentos para tomar diferentes datos en la misma concentración de IPTG en los mismos tiempos e intentar ajustar el modelo con estos datos para ver si se tienen mejores resultados, esto ayudaría a eliminar un poco el ruido que se tiene en las lecturas de fluorescencia y OD.

Con los resultados conseguidos hasta ahora se puede decir que el modelo de Munsky acoplado con crecimiento celular es el modelo más apropiado para recuperar los parámetros de las 96 cepas de E.coli, una vez que se tengan estos conjuntos de parámetros se deben emplear estrategias para agruparlos y compararlos con los grupos obtenidos en el análisis filogenético de las secuencias. Estos análisis filogenéticos pueden ser mejorados, se puede elegir un grupo externo (puede ser la secuencia de una Salmonella) para determinar las races de los árboles.

Esta compararción entre los grupos de parámetros y ramas de los árboles puede ser una primera aproximación para establecer relaciones entre genotipos y fenotipos, dependiendo de los resultados que se obtengan se pueden intentar otras estrategias, puede ser, reduciendo el análisis a las diferentes regiones del operón para identificar cuáles mutaciones están siendo relevantes en los cambios de las constantes, además, se debe implementar el código para establecer relaciones entre las distancias y ver los resultados.

Muchas otras ideas pueden ser exploradas para mejorar el modelo y el análisis entre genotipo y fenotipo. Seguramente surgirán nuevas ideas en la medida que se lleven a cabo las que ya se han propuesto.

4.1. Conclusiones

- El modelo de Munsky acoplado con un crecimiento logístico es el modelo más apropiado para abordar el problema dado que, captura la dinámica del operón y permite recuperar la mayor cantidad de parámetros asociados a regiones de interés.

- Para la solución del problema inverso es necesario usar un término de regularización, en este caso la "sparsity enforcing regularization"

- Es recomendable usar los datos de GFP correspondientes a dos concentración de IPTG, en particular las concentraciones cuatro y cinco, cuatro, dado que corresponden a concentraciones de inductor similar a las que se encontraría la bacteria en condiciones naturales y a la hora de realizar los cálculos ofrece mejor información que otras concentraciones. Adicional a esto, se deben usar los datos de OD para ajustar los parámetros del volumen celular.

- Una primera aproximación para analizar el mapeo genotipo fenotipo es a partir de un análisis filogénetico y un agrupamiento de los parámetros.

4.2. Proyección del Trabajo

Como se ha dicho antes, aquí solo se presentan algunas herramientas para atacar el problema principal que es estudiar la variabilidad de la expresión del operón en una colección de E.coli. Hasta aquí solo se ha presentado un modelo que describe la dinámica, un primer trabajo con las secuencias genéticas y una forma de recuperar los parámetro. Como continuación de esta propuesta queda:

- Recuperar los parámetros para las 96 cepas.
- Implementar un código para el agrupamiento de los parámetros y compararlos con los resultados del análisis filogenético.
- Estudiar las relaciones genotipo fenotipo a partir del código de correlación de distancias.
- Estudiar las regiones del operón de manera individual.

Capítulo 5

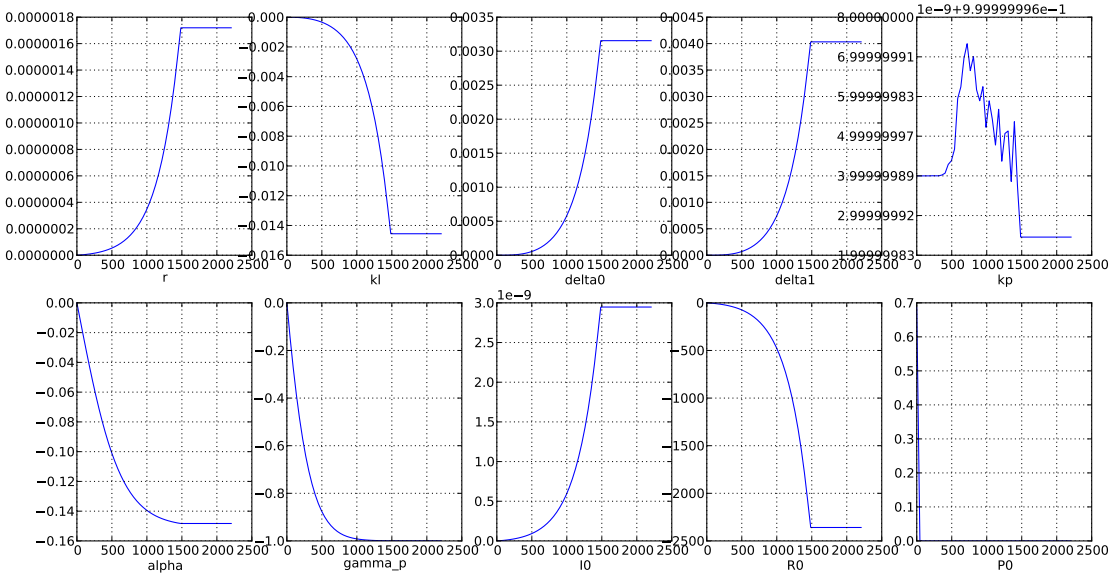
Apéndice

5.1. Resultados de Sensibilidad

Resultados de sensibilidad para las cepas 3, 4 y 5.

Cepa 3.

$$[r : 3{,}09164e^{-01}, k_l : 2e^{-06}, \delta^0 : 1{,}415e^{-03}, \delta^1 : 2{,}32e^{-04},$$
$$k_p : 1{,}99754e^{01}, \alpha : 2{,}430248e^{02}, \gamma_p : 6{,}677e^{-03}, I_0 : 2{,}6354e^{-02},$$
$$R_0 : 9{,}98325e^{01}, P_0 : 1{,}37163e^{-01}]$$

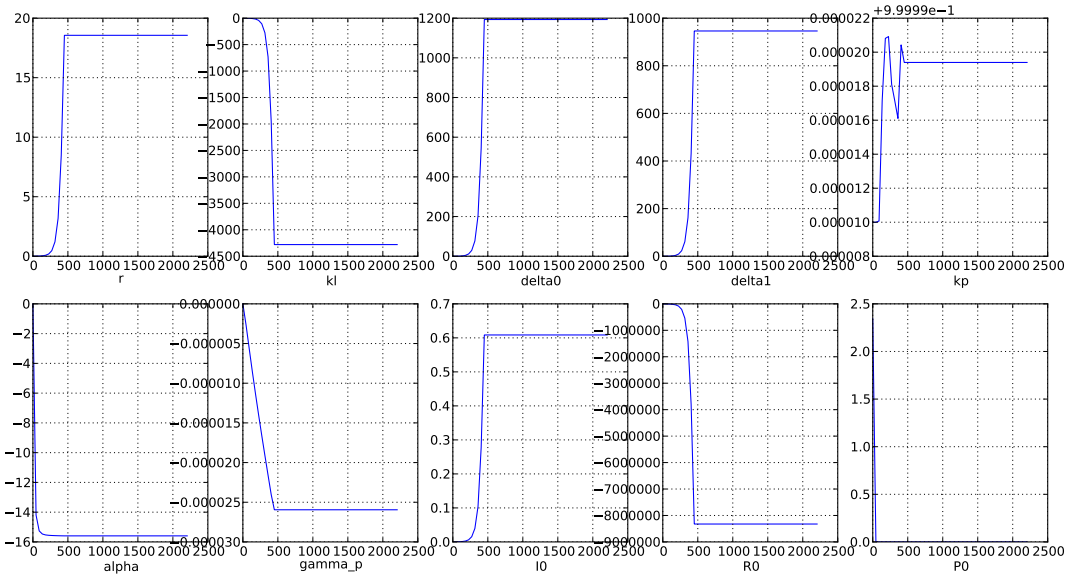


Серия 4.

$$[r : 2{,}87898e^{-01}, k_l : 1{,}185e^{-03}, \delta^0 : 1{,}3232e^{-02}, \delta^1 : 1{,}360e^{-03},$$

$$k_p : 1{,}99877e^{01}, \alpha : 1{,}415137e^{03}, \gamma_p : 1e^{-07}, I_0 : 4{,}58067e^{-01},$$

$$R_0 : 9{,}99808e^{01}, P_0 : 4{,}68378e^{-01}]$$

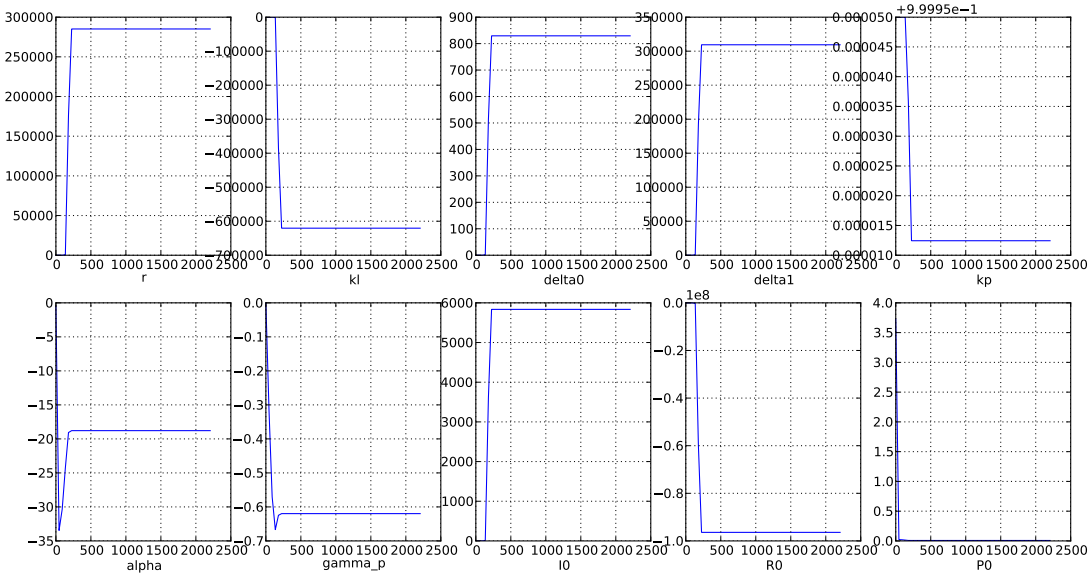


Cepa 5.

$$[r : 3{,}790e^{-03}, k_l : 9{,}508e^{-03}, \delta^0 : 8{,}2e^{-05}, \delta^1 : 3{,}3497e^{-02},$$

$$k_p : 1{,}7831809e^{01}, \alpha : 9{,}7671630e^{03}, \gamma_p : 6{,}849e^{-03}, I_0 : 1{,}9507e^{-02},$$

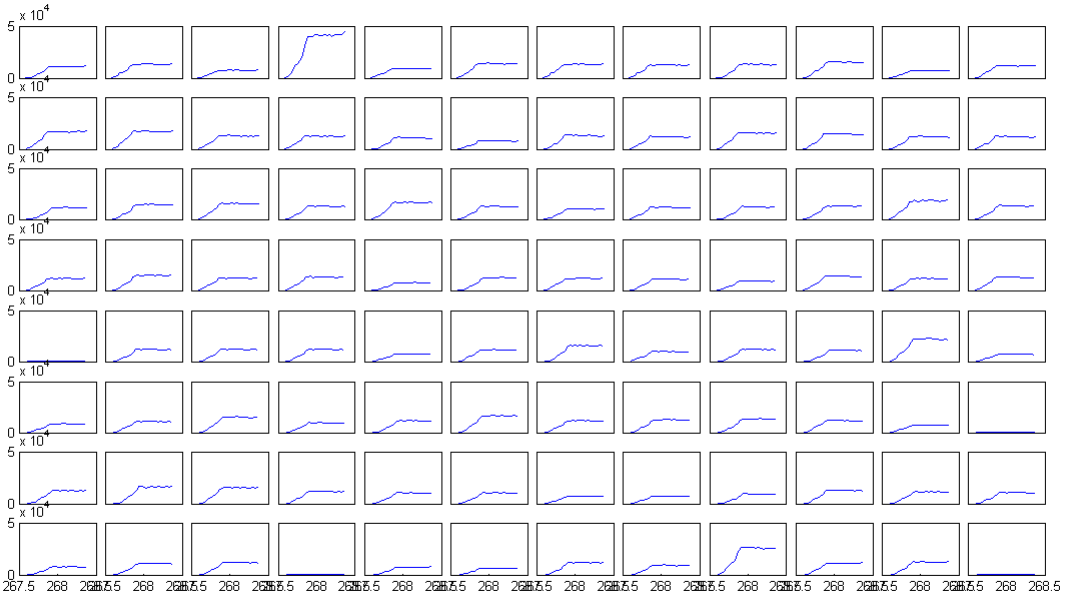
$$R_0 : 6{,}0663618e^{01}, P_0 : 6{,}64578e^{-01}]$$


r
kl
delta0
delta1
kp
alpha
gamma_p
I0
R0
P0
1e8
+9.9995e−1

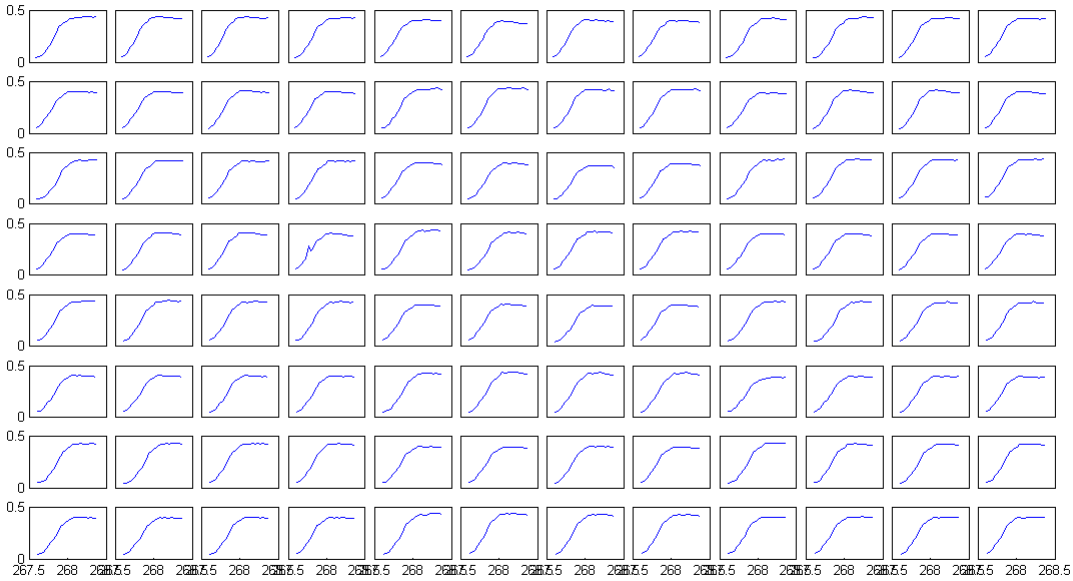
5.2. Datos de la Colección de E.coli

Las siguientes gráficas corresponden a los datos obtenidos para fluorescencia y densidad óptica para las 96 cepas en una concentración de 1mM de IPTG.

Tiempo vs GFP



Tiempo vs OD



Vemos que, aunque todas las células crecieron sin problema, en algunas no se tiene señal de fluorescencia. Por otra parte, no fue posible alinear todas las secuencias de los operones para el análisis filogenético, dado que, algunas de ellas tenían mucho ruido. Así, de un total de 96 cepas el trabajo se verá reducido al estudio de 64. A continuación se presenta una tabla en la que se resume esta información.

A01	A02	-----	----	----	A06	----	----	A09	A10	A11	A12
B01	B02	B03	-----	B05	----	B07	B08	----	B10	B11	B12
C01	C02	-----	C04	----	C06	C07	C08	----	C10	-----	C12
D01	----	D03	D04	----	D06	D07	-----	D09	D10	D11	D12
~~E01~~	E02	E03	E04	E05	----	E07	E08	E09	E10	E11	E12
F01	F02	F03	F04	----	F06	F07	F08	F09	F10	F11	~~F12~~
-----	----	----	G04	----	----	-----	G08	G09	G10	G11	G12
-----	----	H03	~~><~~	H05	----	H07	H08	----	H10	-----	~~><~~

1	9	17	25	33	41	49	57	65	73	81	89
2	10	18	26	34	42	50	58	66	74	82	90
3	11	19	27	35	43	51	59	67	75	83	91
4	12	20	28	36	44	52	60	68	76	84	92
5	13	21	29	37	45	53	61	69	77	85	93
6	14	22	30	38	46	54	62	70	78	86	94
7	15	23	31	39	47	55	63	71	79	87	95
8	16	24	32	40	48	56	64	72	80	88	96

Rosa: Falta de secuencia

Rojo: Sin señal de GFP

La tabla simula una placa con las 96 cepas, las líneas punteadas indican que no fue posible alinear esa secuencia y las cruces que esa cepa no presenta señal de fluorescencia.

5.3. Análisis de identificabilidad y sensibilidad

A continuación se presenta el análisis de sensibilidad e identificabilidad para el modelo 1 y el modelo 3 con algunas modificaciones. Se tomo el modelo tres, que corresponde al modelo que se propuso y se le hicierón algunas modificaciones tales como: 1) Se agregó una ecuación adicional para el ingreso de IPTG, se elimin el crecimiento celular y se hizo el ajuste con los datos multiplicando el valor de la proteína por la cantidad de células en ese momento tal como se hace con el modelo de munsky, pero los resultados no son mejores que los que se obtienen con el modelo de Munsky.

5.4. Modelo 1 - Guido

score cepa1 $= 9{,}8370e+03$

$$\frac{dM}{dt} = k_M - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$

5.4.1. Identificabilidad de Parámetros

Dados los vectores:

$$\hat{p} = [\hat{k_m}, \hat{\gamma_M}, \hat{k_p}, \hat{\gamma_p}, \hat{M}_0, \hat{P}_0]$$

$$p = [k_M, \gamma_M, k_p, \gamma_p, M_0, P_0]$$

Calculamos las derivadas correspondientes para aplicar el teorema de series.

$$
\begin{aligned}
a_0(p) &= y_m(t_0, p) = P_0 \\
a_1(p) &= P' = k_p M_0 - \gamma_p P_0 \\
a_2(p) &= P'' = k_p M' - \gamma_p P' \\
&= k_p(k_M - \gamma_M M_0) - \gamma_p(k_p M_0 - \gamma_p P_0) \\
&= k_p k_M - k_p \gamma_M M_0 - \gamma_p k_p M_0 - \gamma_p^2 P_0 \\
a_3(p) &= -k_p \gamma_M M' - \gamma_p k_p M' - \gamma_p^2 P' \\
&= -k_p k_M \gamma_M + k_p {\gamma_M}^2 M_0 - \gamma_p k_p k_M + \gamma_p k_p \gamma_M M_0 - {\gamma_p}^2 k_p M_0 + {\gamma_p}^3 P_0
\end{aligned}
$$

Debe cumplirse que:

$$
\begin{aligned}
a_0(\hat{p}) &= a_0(p) \\
a_1(\hat{p}) &= a_1(p) \\
&\vdots
\end{aligned}
$$

Igualando los términos se tiene que γ_M,γ_p,P_0 y los productos $k_p k_M$, $k_p M_0$ son identificables. Son necesarias condiciones adicionales para conseguir la identificabilidad de todos los parámetros

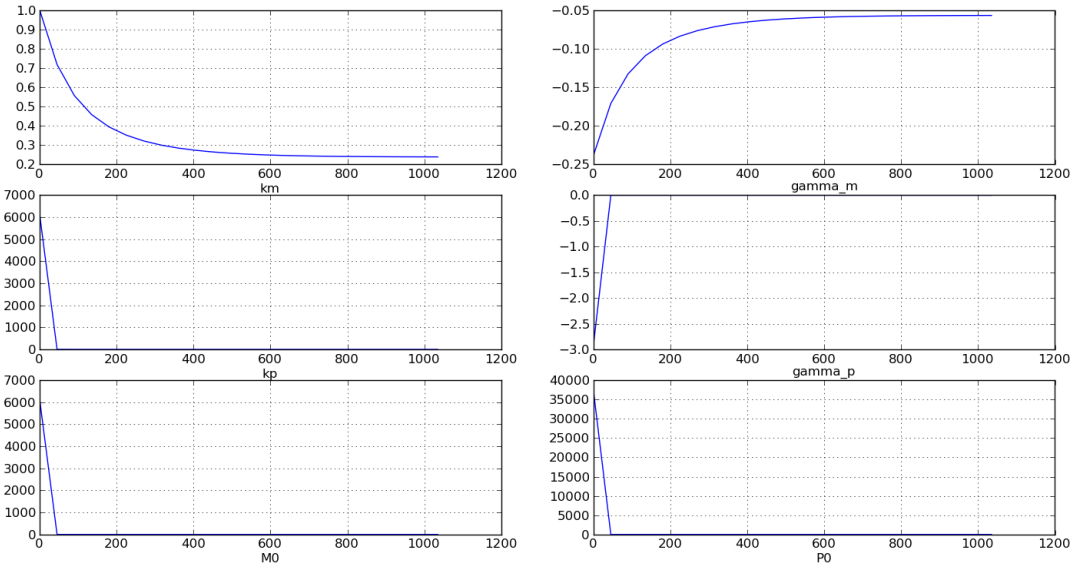
5.4.2. Análisis de Sensibilidad

La ecuación de sensibilidad a resolver es

$$
\dot{S}_i = \begin{bmatrix} -\gamma_M & 0 \\ k_p & -\gamma_p \end{bmatrix} s_i + \frac{\partial f}{\partial p_i}
$$

donde

$$\frac{\partial f}{\partial k_M} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial \gamma_m} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial k_p} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \frac{\partial f}{\partial \gamma_p} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$



5.5. Modelo

score con 5000 iteraciones = 1.2325e+04

Este modelo corresponde al que hemos propuesto como modelo 3 con algunas modificaciones, se ha incluido una ecuación adicional para el ingreso de IPTG, se elimina el término del volumen en el modelo y se multiplica el total de GFP por V como se hace con el modelo de Musky.

$$\frac{dM}{dt} = k_M N \frac{1 + k_I I^2}{(1 + k_R R_T + k_I I^2)} - \gamma_M M$$
$$\frac{dP}{dt} = k_p M - \gamma_p P$$
$$\frac{dI}{dt} = r(I_{max} - I)$$

5.5.1. Identificabilidad de Parámetros

El teorema de series deja de ser práctico en este caso dada la dimensión de sistema, los cálculos se vuelven casi imposibles aún haciendo uso de herramientas computacionales.

5.5.2. Análisis de Sensibilidad

La ecuación de sensibilidad a resolver es

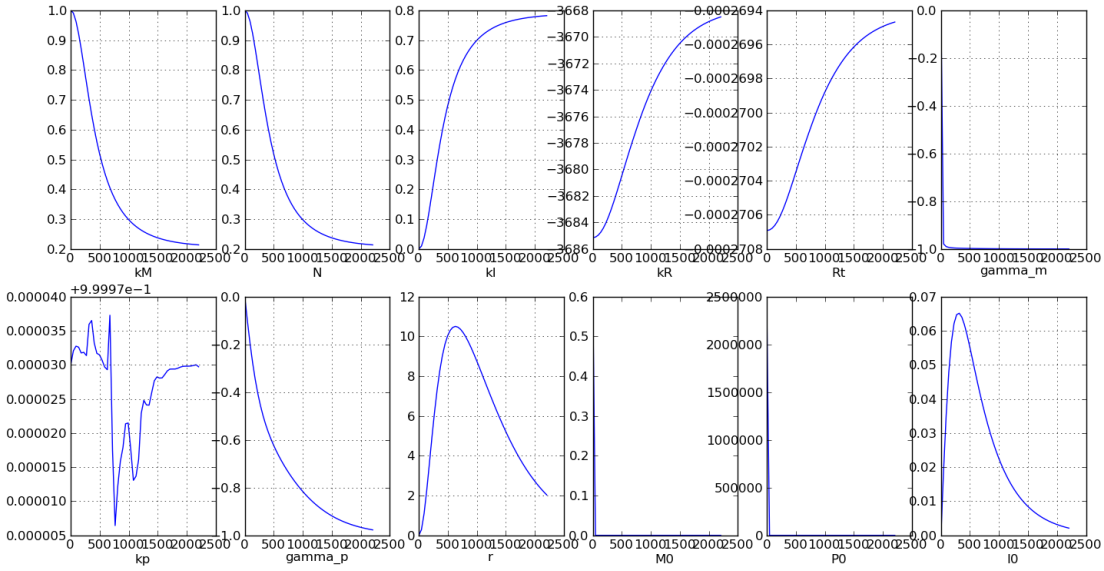
$$\dot{S}_i = \begin{bmatrix} -\gamma_M & 0 & \frac{2k_M N k_I R_T I}{(1+k_R R_T + k_I I^2)^2} \\ k_p & -\gamma_p & 0 \\ 0 & 0 & -r \end{bmatrix} s_i + \frac{\partial f}{\partial p_i}$$

donde

$$\frac{\partial f}{\partial k_M} = \begin{bmatrix} \frac{N(1+k_I I^2)}{1+k_R R_T+k_I I^2} \\ 0 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial N} = \begin{bmatrix} \frac{k_M(1+k_I I^2)}{1+k_R R_T+k_I I^2} \\ 0 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial k_I} = \begin{bmatrix} \frac{k_M N k_R R_T I^2}{(1+k_R R_T+k_I I^2)^2} \\ 0 \\ 0 \end{bmatrix}$$

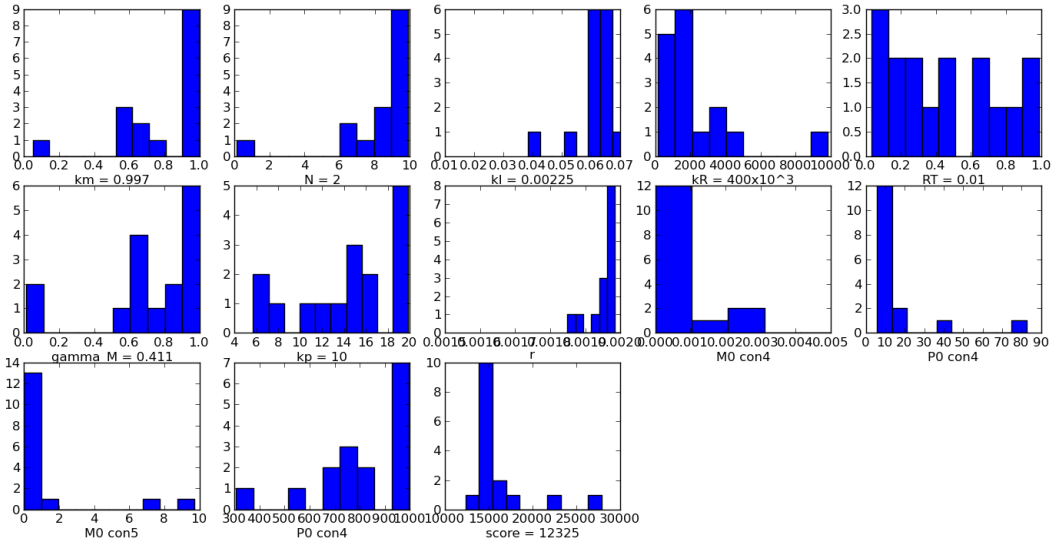
$$\frac{\partial f}{\partial k_R} = \begin{bmatrix} \frac{-R_T k_M N(1+k_I I^2)}{(1+k_R R_T+k_I I^2)^2} \\ 0 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial k_R} = \begin{bmatrix} \frac{-k_R k_M N(1+k_I I^2)}{(1+k_R R_T+k_I I^2)^2} \\ 0 \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial \gamma_m} = \begin{bmatrix} -M \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial f}{\partial k_p} = \begin{bmatrix} 0 \\ M \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial \gamma_p} = \begin{bmatrix} 0 \\ -P \\ 0 \end{bmatrix} \quad \frac{\partial f}{\partial k_R} = \begin{bmatrix} 0 \\ 0 \\ I_{max} - I \end{bmatrix}$$



Distribución de los Parámetros

5000 iteraciones



Los resultados no mejoran con esta estrategia, al contrario, presentan una mayor desviación estandar.

5.6. Código Implementado para la Estimación de Parámetros

5.6.1. Código

```
1. Inicializacion de las variables:

   Tiempo de las medidas
   Datos de GFP
   Datos de OD
   Degradacin de la protena
   Concentraciones de IPTG

2. Definicion de las Ecuaciones

    def eqs1(x, t, p):
                f = zeros([2])
                iptg_in = self.Imax1*(1.0 - exp(-p[0]*t))
                f[0] = p[1] - (p[2] + p[3]*iptg_in)*x[0]
                f[1] = p[4]/(1.0 + p[5]*x[0]**2) - self.d_p*x[1]
                return f

def eqs2(x, t, p):
                f = zeros([2])
                iptg_in = self.Imax2*(1.0 - exp(-p[0]*t))
                f[0] = p[1] - (p[2] + p[3]*iptg_in)*x[0]
                f[1] = p[4]/(1.0 + p[5]*x[0]**2) - self.d_p*x[1]
                return f

3. Solucion del Sistema de Ecuciones Diferenciales

    def evaluate1(p):
            y0 = [p[6],p[7]]
                ySoln1 = integrate.odeint(self.eqs1, y0, self.tdata, args=(p,))
                return ySoln1

def evaluate2(p):
            y0 = [p[8],p[9]]
                ySoln2 = integrate.odeint(self.eqs2, y0, self.tdata, args=(p,))
                return ySoln2

4. Residuos: Estos corresponden a los terminos
            ||F(p)*V - datosGFP||  y  ||a*V - datosOD||

    def residue1(self,p):
                ySoln1 = self.evaluate1(p)
                vol = p[10]*p[11]/(p[11] + (p[10] - p[11])*exp(-p[12]*self.tdata))
        res1a = vol*ySoln1[:,1] - self.GFPcepa1[:,self.myindex1]
                res1b = p[16]*vol - self.ODcepa1[:,self.myindex1]
```

```
                return res1a, res1b

def residue2(self,p):
                ySoln2 = self.evaluate2(p)
                vol = p[13]*p[14]/(p[14] + (p[13] - p[14])*exp(-p[15]*self.tdata))
        res2a = vol*ySoln2[:,1] - self.GFPcepa1[:,self.myindex2]
                res2b = p[17]*vol - self.ODcepa1[:,self.myindex2]
                return res2a, res2b
```

4. Corresponde al termino que se quiere Optimizar

```
||F_1(p)*V_1 -datosGFP_1|| + ||F_2(p)*V_2 - datosGFP_2||+||a_1*V_1 - datosOD_1|| + ||a_2*V_2 - datosOD_2|| + R(p)

    def error(self,p):
                try:
                  reg_term = p[18]*linalg.basic.norm(p[:-1],ord=1)
                  res1a,res1b = self.residue1(p)
                  res2a,res2b = self.residue2(p)
                  abserror1 = linalg.basic.norm(res1a)**2
                  abserror2 = linalg.basic.norm(res1b)**2
                  abserror3 = linalg.basic.norm(res2a)**2
                  abserror4 = linalg.basic.norm(res2b)**2
                  abserror = abserror1 + abserror2 + abserror3 + abserror4 + reg_term
                  except ValueError:
                  abserror = 10.0**10
                  return abserror
```

5. Grafica la Solucion

```
def show_soln(self,p):
ySoln1 = self.evaluate1(p)
vol = p[13]*p[14]/(p[14] + (p[13] - p[14])*exp(-p[15]*self.tdata))
subplot(211)
plot(self.tdata,ySoln1[:,1]*vol,'r.-')
plot(self.tdata,self.GFPcepa1[:,self.myindex1],'k.-')
grid()
ySoln2 = self.evaluate2(p)
vol = p[13]*p[14]/(p[14] + (p[13] - p[14])*exp(-p[15]*self.tdata))
subplot(212)
plot(self.tdata,ySoln2[:,1]*vol,'r.-')
plot(self.tdata,self.GFPcepa1[:,self.myindex2],'k.-')
grid()
```

6. Define los intervalos en que viven los parametros

```
bounds = [(0.0,1.0), (0.0,1.0), (0.0,1.0), (0.0,1.0), (0.0,20.0),(0.0,20.0),
          (0.0,1000.0), (0.0,1000.0),
          (0.0,1000.0), (0.0,1000.0),
          (0.0,10000.0), (0.0,100.0), (0.0,1.0),
          (0.0,10000.0), (0.0,100.0), (0.0,1.0),
          (0.0,1.0), (0.0,1.0), (0.0,1.0)]
```

```
7. Para la optimizacion se utilizan datos correspondientes a dos de las concentraciones medidas.

ecoli = ForwardMapping(myindex1 = 4, myindex2 = 5, cepa = 1)

8. Resuelve el problema de optimizacion utilizando Evolucion Diferencial

solver = my_desolver.DESolver(ecoli.error, bounds, 500, 10000,
                              method = my_desolver.DE_RAND_1,
                              scale=[0.1,0.9],
                              crossover_prob=0.9,
                              goal_error=1e-5, polish=False, verbose=True,
                              use_pp = False, pp_modules=['numpy'])
```

Todos los códigos fueron implementados en Python.

Bibliografía

[1] Acero, S. (2001). *Algoritmo para Análisis Fiologenético: UPGMA* Tesis: Universidad del Norte-Barranquilla

[2] Arribere, P. (2007).*Ingeniera metablica de la gluclisis mediante tcnicas de simplificacin de modelos dinmicos no lineales*. Proyecto fin de carrera. Universidad de Málaga.

[3] Arriola. L., Hyman. J.(2007) *Being Sensitive to Uncertainy*. Computing in Science & Engineering.

[4] Barbosa, L. O: *Persistencia, extincin e invasin de una epidemia: Retroalimentacin de un modelo matemtico*. Tesis.

[5] Bloom, G. and P. W. Sherman (2005). *"Dairying barriers affect the distribution of lactose malabsorption."* Evolution and Human Behavior 26(4): 301-312.

[6] Borneman, A. R., T. A. Gianoulis, et al. (2007). *"Divergence of transcription factor binding sites across related yeast species."* Science 317(5839): 815-819.

[7] Brem, R. B. and L. Kruglyak (2005). *"The landscape of genetic complexity across 5,700 gene expression traits in yeast"* Proceedings of the National Academy of Sciences of the United States of America 102(5): 1572-1577.

[8] Brem, R. B., J. D. Storey, et al. (2005). *"Genetic interactions between polymorphisms that affect gene expression in yeast"* Nature 436(7051): 701-703.

[9] Brown, K. M., C. R. Landry, et al. (2008). *Çascading transcriptional effects of a naturally occurring frameshift mutation in Saccharomyces cerevisiae."* Molecular Ecology 17(12): 2985-2997.

[10] Capistrn, M.A., Moreles, M.A, Lara, Bruno (2009) *Parameter Estimation of some Epidemic Models.The Case of Recurrent Epidemics Caused by Respiratory Syncytial Viru.* Bulletin of Mathematical Biology 71: 1890-1901

[11] Chen, W.Y., Bokka, S (2005) *Stochastic modeling of nonlinear epidemiology.* Journal of Theoretical Biology 234: 455-470.

[12] Ching, A., K. Caldwell, et al. (2002). *"SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines."* BMC genetics 3(1): 19.

[13] Collins, J. J., Isaacs, F.J., Hasty, J. and Cantor, C. R.(2003) *Prediction and measurement of an autoregulatory genetic module.* PNAS. Vol 100. N 13: 7714-7719.

[14] Collins, J.J. , Guido, N. J., Wang, Xiao., Adalsteinsson, McMillen, D., Hasty, J. and Cantor, C.R.(2006) *A bottom-up approach to gene regulation.* Nature. Vol: 439: 856-860.2

[15] Cooper, T. F., D. E. Rozen, et al. (2003). *"Parallel changes in qene expression after 20,000 generations of evolution in Escherichia coli."*Proceedings of the National Academy of Sciences of the United States of America 100(3): 1072-1077.

[16] Cooper, T. F., S. K. Remold, et al. (2008). *.Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli."* Plos Genetics 4(2).

[17] Engl. H et al (2009) *Inverse Problem in systems biology.* IOP Publishing ltd. 25:1-55

[18] Fan, J.-B., X. Chen, et al. (2000). *"Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays."* Genome Research 10(6): 853-860.

[19] Fay, J. C., H. L. McCullough, et al. (2004). *"Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae."* Genome Biology 5(4).

[20] Gross, C.A, Von Hippel, P.H,Recizin, A. and Wang, A.(1974) *Non-specific DNA Binding of Genome Regulating Proteins as a Biological Control Mechanism: 1. The lac Operon: Equilibrium Aspects*Proc. Nat. Acad. Sci. USA, Vol. 71, No. 12, pp. 4808-4812

[21] Hemberg, M and Barahona, M (2007) *Perfect Sampling of the Master Equation for Gene Regulatory Network.* Biophysical Journal. 93: 401-410

[22] Hengl,S., Kreutz, C., Timmer, J. and Maiwald T. (2007). *Data-based iden-*

tifiability analysis of non-linear dynamical models. Bioinformatics, Vol. 23 no. 19 2007, pages 2612-2618.

[23] Hindmarsh. A, Serban. R.(2009) *User Documentation for cvodes v2.6.0. Center for Applied Scientific Computing Lawrence Livermore National Laboratory*.

[24] Kaern, M., Elston, T. C., Blake, W. J. and Collins, J.J.(2005) *Stochasticity in gene expression: from theories to phenotypes*. Nature. Vol 6: 451-464

[25] Keener, J. Sneyd, J.(1998) *Mathematical Physiology* Springer, Vol 8,p 432-438

[26] Krakauer et al (2011). *The challengesandscopeoftheoreticalbiology*. Journal of Theoretical Biology 276. 269-276

[27] Kuhlman, T., Zhang, Z., Saier, M. H. and Hwa, T (2007) *Combinatorial transcriptional control of the lactose operon of Escherichia coli*. PNAS. Vol 104, N14: 6043-6048

[28] Landry, C. R., P. J. Wittkopp, et al. (2005). *Çompensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila."* Genetics 171(4): 1813-1822.

[29] Legendre, P. and Legendre, L. (1998). *Numerical Ecology* ELSEVIER

[30] Lemos, B., C. D. Meiklejohn, et al. (2005). *Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories."* Evolution 59(1): 126-137.

[31] Lestas, I, et al (2008) *Noise in Gene Regulatory Networks*. Special Issiu on Systems biology. 189:200

[32] lto, Y., Toyota, H., Kaneko, K., Yomo, T(2009) *How selection affects phenotypic fluctuation*. Molecular system biology.5:1,7

[33] Miao.H. et al: *On Identificability of nonlinear ODE models and applications in viral dynamics*

[34] Moles y Mendel (2011). *Global Optimization Methods Parameter Estimation in Biochemical Pathways: A Comparison of*. Genome Research.

[35] Munsky, B., Trinh, B. and Khammash, M.(2009) *Listening to the noise: random fluctuations reveal gene network parameters*. Molecular System Biology

[36] Ozbudak, E. M., Thattai, H. N, Lim, B. I, Shariman, and A. van Oudenaarden(2004) *Multistability in the lactose utilization network of Escherichia coli*. Nature. 427: 737-740

[37] Ponciano. J., Capistrn. M. (2011) *First principles modeling of nonlinear incidence rates in seasonal epidemics*.PLOS 7: 1-14

[38] Prozano, W. E. (1997) *Identification of Parametric Models from experimental data*. Great Britain. Springer

[39] Rockman, M. V. and L. Kruglyak (2006). *"Genetics of global gene expression."* Nature Reviews Genetics 7(11): 862-872.

[40] S. B. Lee. Bailey J. E (1984) *Genetically Structured Models for lac*

Promoter-Operator Function in the Chromosome and in Multicopy Plasmids: lac Promoter Function. Biotechnology and Bioengineering, VOL. 26

[41] Santilln, M(2008) *Bistable Behavior in a Model of the Lac Operon in Escherichia coli with Variable Growth Rate*. Biophysical Journal Vol 94:2065-2081

[42] Santilln, M., Mackey, M.C. and Zeron, E.S.(2007) *Origin of Bistability in the lac Operon*. Biophysical Journal. Vol: 92: 3830-3842

[43] Stamatakis ,M. Mantzaris N. V (2009) *Comparison of Deterministic and Stochastic Models of the lac Operon Genetic Network*. Biophysical Journal 96: 887-906

[44] Talmud, P. J. (2007). *"Gene-environment interaction and its impact on coronary heart disease risk."* Nutrition Metabolism and Cardiovascular Diseases 17(2): 148-152.

[45] Tirosh, I., A. Weinberger, et al. (2008)..*On the relation between promoter divergence and gene expression evolution."* Molecular Systems Biology 4.

[46] Uri Alon, Kalisky, T.(2007) *Cost-benefit theory and optimal design of gene regulation functions* Phisical Biology, 229-245.

[47] Vilar, J. M.G, Guet, C.C and Leibler, S(2003) *Modeling network dynamics: the lac operon, a case study*. Journal of Cell Biology. Vol 161, N3: 471-476

[48] Vogel, Curtis. (2002). *Computational Methods for Inverse Problems* Society for Industrial and Applied Mathematics Philadelphia.

[49] Wong, P., Gladney, S., Keasling, J.D (1997) *Mathematical Model of the lac Operon: Inducer Exclusion,Catabolite Repression, and Diauxic Growth on Glucose and Lactose*Biotechnol. Prog.,Vol. 13, No. 2

[50] Yao, K.Z. et al (2003). *Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimablity and Experimental Design* Polymer Reaction Engineering. Vol 11, N3. pp. 563-588

[51] Yildirim, N., and M. C. Mackey (2003) *Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data*. Biophys. J. 84: 2841-2851

[52] Yue, P. and J. Moult (2006).*Ïdentification and Analysis of Deleterious Human SNPs."* Journal of Molecular Biology 356(5): 1263-1274.

[53] Yusuke, T. M. and Sano, M (2006) *Regulatory Dynamics of Synthetic Gene Networks with Positive Feedback*. JMB. 359: 1107-1124

[54] Zeron, E. S. and Santilln, M (2010) *Distributions for negative - feedback - regulated stochastic gene expression: Dimension reduction and numerical solution of the chemical master equation*. Journal of Theoretical Biology. 264: 377-385

[55] Zhu, J., B. Zhang, et al. (2008). *Ïntegrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks."* Nature Genetics 40(7): 854- 861.

[56] Zielinski, K., Rainer, L. (2008) *Stopping Criteria for Differential Evolution in Constrained Single-Objective Optimization*. Institute for Electromagnetic Theory and Microelectronics, University of Bremen.