Soluciones analíticas a sistemas de expresión de genes con feedback negativo.

Juan Carlos Linares Rugeles

Tutor: Dr. Juan Manuel Pedraza

Universidad de los Andes

23 de junio de 2016

.

En la misma manera, el mundo no es la suma de todas las cosas que están en él. Es la red de conexiones infinítamente compleja entre estas. Como en el significado de las palabras, las cosas toman su significado sólo en relación con las demás.

La invención de la soledad, Paul Auster

.

Índice

1. Introducción

1.1. Motivación

El análisis de sistemas de expresión de genes en organismos unicelulares es un tema de importancia, puesto que es la via de estudio teórico que se tiene del funcionamiento interno de estos organismos. Se generan modelos computacionales y predicciones en multiples laboratorios a lo largo de todo el globo terráqueo en base a estos sistemas. Sin embargo, aparte de la experimentación y el modelamiento computacional es un área de estudio que se ha visto un poco trancada por la falta de formalización matemática y desarrollo teórico cuantificable. Aquí es donde se hace útil es uso de métodos de cuantificación y de estudio de procesos aleatorios como por ejemplo los propuestos por la teoría de la información y el método de la ecuación maestra.

Teniendo en cuenta lo anterior nace un interés por idear en un estudio de la manera más simple posible, aún así usando las herramientas de cuantificación que nos brinda la teoría de la información propuesta por Shannon (30 de abril de 1916 – 24 de febrero de 2001); en donde se puedan comparar resultados teóricos con prácticos en tiempos relativamente cortos. En respuesta a esto nace el enfoque que se le ha dado a la teoría de la información en torno al flujo de información en organismos unicelulares.

Dada la problemática expuesta, el objetivo de este trabajo es resolver de manera analítica un sistema genérico en abstracto, en donde un compuesto interactúe consigo mismo, ya sea reprimiendose o cooperando. Lo importante de este estudio es que fue planeado y llevado a cabo, con la grán ayuda de mi director de tesis Juan M. Pedraza, con el fin de que no se asumiera ningún tipo particular de ruido en el sistema; sólo hacer el modelamiento con base en el método de la ecuación maestra para dicho sistema (y jugando un poco con este a conveniencia).

1.2. Recuento Histórico

Desde los inicios de la teoría de la información con los estudios de Shannon [[1],[2]], se ha valorado la elegancia de sus ideas para cuantificar algo tan abstracto y tan cotidiano como la información. De las primeras cosas que suscito la visión de Shannon fue la curiosidad por cómo podría estudiarse el ser

humano desde esta perspectiva. Primero se intaron llegar a interpretaciónes de la teoría de la información un poco más salidas del ámbito de la comunicación [3] (tema de interés en las publicaciones de Shannon nombradas). Ideas como que este nuevo campo de estudio que había creado/descubierto Shannon proveía un criterio para medir la organización de ciertas variables; en resumen, adecuaciónes y discusiones sobre los conceptos básicos. Siguiendo otro trabajo de Shannon en donde estudia la redundancia en el inglés impreso, el psicologo Fred Attneave publica un trabajo acerca de la posible aplicación de la teoría de la información a la percepción visual [4] en el año 1954. En este trabajo trata situaciones experimentales como la de mirar la distribución de puntos que se generaba al pedir a sujetos resumir una forma bidimensional en puntos con el objetivo de mostrar cierta redundancia en el hecho de exhibir la figura completa, o desde mi punto de vista, mostrar cúal era la compresión más común de la figura.

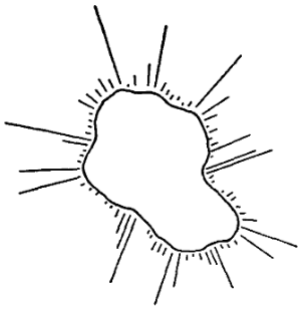

Figura 1: Sujetos intentaban aproximar la figura cerrada en 10 puntos. Las barras que salen de esta indican la frecuencia con la que se escogieron puntos en dicho punto de la figura. Tomado de [4]

Este tipo de experimentos fomentaron aún más curiosidad, logrando incluso nuevos acercamientos como el propuesto por el neurocientífico Horace Barlow en 1961 [5]. Fue uno de los primeros en proponer estudios sobre el flujo de información en núcleos (o células) de reelevo sensorial en el tálamo, presentando como hipótesis que estas celulas funcionan como puntos de control donde el flujo de información es modulado, es decir, que funcionaban como filtro para la información que es en efecto relevante (reducción de la redundancia). Barlow habla sobre la codificación de los mensajes y la capa-

cidad de un camino nervioso en función de la información promedio de los mensajes que se pueden recibir (entropía del conjunto) y la duración promedio de los mensajes. A partir de esto se generaron con el tiempo estudios cada vez más sofisticados y cuantitativos en neurociencia y el flujo de la información a través de redes neuronales y sistemas nerviosos [[6], [7]]. Tanto así que rápidamente este estudio se fue ligando con la biofísica teórica, lógica de sistemas biológicos, biología cuántica, aspectos de termodinámica y mecánica estadistica [[8], [9], [10]]. Nunca dejando de lado a la teoría de la información, sino más bien permeándolos de esta.

Intentando cuantificar y analizar de manera más exacta la transmisión de información en redes neuronales, los estudios se fueron enfocando al flujo de información en organismos unicelulares, en donde ahora sí jugó un papel fundamental la bioquímica y los efectos termodinámicos (que se creía podían llegar a jugar el papel de ruido en los canales de información). En 2006 y 2007 se proponen estudios teóricos sobre procesamiento de señales en redes bioquímicas pequeñas, tales como la expresión de genes[[11], [12], [13]].

Se ha trabajado también en modelos en donde ligandos se enlazan, independientemente a los otros, a múltiples sitios (como por ejemplo, sitios específicos a lo largo del ADN), teniendo en cuenta que la energía de ligadura a los sitios se encuentra determinada con un estado del sistema en específico [[14], [15]]. En estos modelos se ha concluido, por medio de un análisis en el ruido del sistema, que la cooperatividad entre las interacciones en los entrelazamientos de las múltiples moléculas de señalización ayuda a llegar más fácilmente a los límites físicos de la quimiorrecepción, sin necesidad de disminuir el ruido. Así, juntando conceptos como cooperatividad, canales de información y ruido asociado a estos, se han hecho intentos por formalizar la noción de transmisión de información en sistemas cada vez más complejos y físicamente realistas [16].

2. Marco Teórico

2.1. Formalización de conceptos

Como vimos anteriormente, el campo de la teoría de la información aplicada a organismos unicelulares es un campo de investigación emergente, y como tal posee ciertos problemas. Problemas por ejemplo de modelación que poco a poco se van resolviendo con ideas innovadoras como la que proponen Taylor, Tishby y Bialek en su trabajo del 2007 *Information and fitness* [17]. En este trabajo se propone tratar el sistema bacteria-entorno por medio de la asignación de un vector de caracteristicas a cada componente del sistema. Las variables internas del sistema (las de la bacteria) representadas por $\vec{g}$, con componentes como la tasa de expresión promedio de ciertas proteinas, y las externas (del entorno) por $\vec{s}$, con componentes como la concentración de ciertos compuestos químicos en el entorno, termperatura promedio, etc.

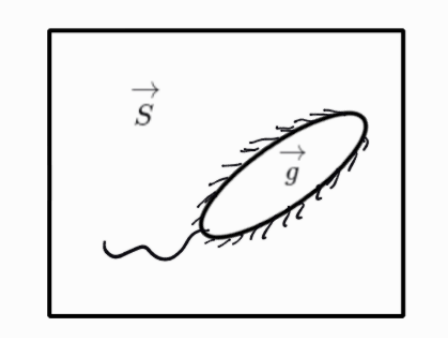

Figura 2: Representación de una bacteria Escherichia coli (*E.coli*), con vector de variables internas $\vec{g}$ en un ambiente de variables externas $\vec{s}$.

Se proponen estas formas compactas de nombrar las variables del sistema con el fin de facilitar la tarea de hallar una relación entre la información que un organismo tiene sobre su entorno y su fitness. La funcion de fitness la podemos describir entonces como función de ambos tipos de variables $f \equiv f(\vec{g}, \vec{s})$. La idea detras de este artículo es pensar que podemos fabricar esta función, ya sea estimandola a partir de datos experimentales (caso que se expondrá luego), o de manera teórica a partir de variables claves del sistema, como lo proponen en [17].

Intuitivamente podríamos pensar que tal funcion de fitness $f(\vec{g}, \vec{s})$ sería más grande a medida que la información mutua entre $\vec{g}$ y $\vec{s}$ es mayor. Es decir, que entre más correlacionadas estén las dos variables, entre más determine de manera certera una a la otra, mejor estará adaptado el organismo al ambiente. En general esto no pasa, puesto que el proceso de lectura del ambiente no es perfecto por ruidos térmicos intrínsecos en los procesos involucrados. Lo que sí es seguro es que el proceso de censado sería más exacto si el hecho de censar el medio ambiente fuera gratuito en términos de gastos energéticos. Por ejemplo si estamos hablando de una bacteria, esta podría estar gastando la energía que está empleando en censar el medio en reproducirse, lo cual aumentaría su fitness. Luego entonces ¿por qué no simplemente nunca censar el ambiente y reproducirse indefinidamente? De hecho se comprobó experimentalmente que también es una estrategia válida, sin embargo se ha demostrado [18] que lo mejor es una colaboración de ambos métodos, ya que al reproducirse sin censar se corre peligro de que las siguientes generaciones no sobrevivan al ambiente si este es cambiante. En todo caso, dado lo anterior podemos afirmar que la función de fitness tendrá un máximo para un cierto vector de variables internas $\vec{g}$ dado uno de externas $\vec{s}$, pues para todo ambiente existirá un cierto fenotipo del organismo que presente una ventaja sobre los otros, y como el óptimo de un ambiente no es necesariamente el mismo que el de otro, el máximo de f también cambiará con $\vec{s}$.

Dado que la función de fitness puede llegar a ser muy complicada, se intenta siempre tratar con su promedio estadístico sobre todo el sistema:

$$\langle f \rangle = \int d^K s \int d^D g P(\vec{g}, \vec{s}) f(\vec{g}, \vec{s}) \tag{1}$$

Ahora para tener un fitness óptimo, que no sea redundante respecto a la información que adquiere la bacteria del ambiente, se puede escoger entre todas las distribuciones condicionales que lleven al mismo $< f >$ y minimizar entre estas la información mutua entre $\vec{g}$ y $\vec{s}$:

$$I(\vec{g}; \vec{s}) = \int d^K s \int d^D g P(\vec{g}, \vec{s}) log_2 \left[\frac{P(\vec{g}|\vec{s})}{P(\vec{g})} \right] bits \tag{2}$$

La existencia entonces de un proceso (muy probablemente poco sencillo) por el cuál podemos llegar a la información mínima dado un $\langle f \rangle$, es decir $I_{min}(\langle f \rangle)$, nos dice que al estudiar el fenotipo de cierto organismo sometido a diferentes ambientes se encontrará que este provee un mínimo de información acerca del ambiente. En otras palabras, se logra mostrar que para mantener una tasa de crecimiento promedio sobre un conjunto de condiciones se debe llevar siempre una representación interna del entorno capturada en un número mínimo de bits. De la misma estructura de la función de información podemos ver que a medida que se aumenta $\langle f \rangle$ también se aumenta $I_{min}(\langle f \rangle)$, lo cual define una noción de organismo óptimo y una dirección evolutiva que promueve el aumento de la capacidad de recopilar información (no redundante) del medio.

En aras de mostrar que es posible hallar la función de fitness de manera experimental podemos referenciarnos al experimento con en *E.coli* propuesto por Dekel y Alon [19], en donde se estima la función de fitness como función de la concentración externa de lactosa (s) y los niveles de expresión de operón lac; compuesto (g) requerido para el metabolismo de la lactosa en *E.coli* (ver figura 3).

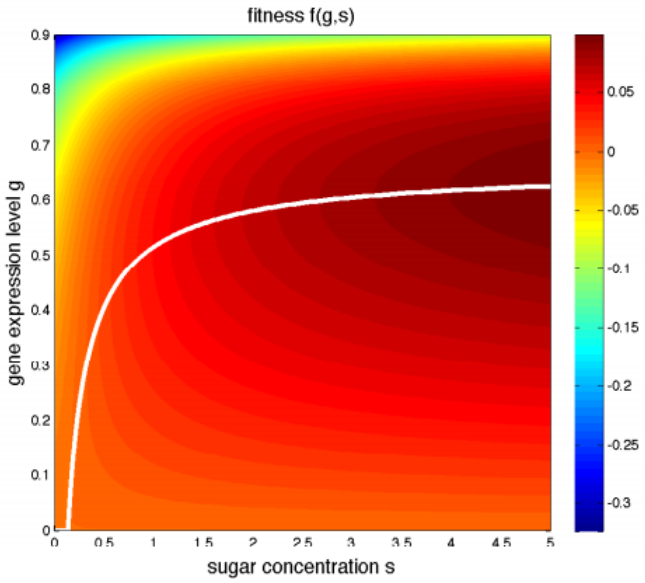

Figura 3: Tasa de crecimiento de *E.coli* como función del medio (concentración de lactosa) y la expresión del gen lacZ. La medida del fitness se muestra normada por la tasa de crecimiento cuando ambas condiciones eran cero. Las concentraciones de lactosa son medidas de tal manera que el beneficio máximo está en s=1, y los niveles de expresión del gen en unidades normadas por lo que la celula puede mantener. La linea blanca traza el nivel de expresión optimo para cada valor de s. Los datos provienen del experimento de Dekel y Alon [19], sin embargo la imagen fue construida por Taylor, Tishby y Bialek [17]

2.2. Ecuación maestra

Uno de los protagonistas principales de esta historia es el ruido, entendido como las fluctuaciones presentes en los procesos intracelulares, y si deseamos estudiar lo más realista posible el sistema propuesto debemos tener en cuenta el ruido. Se piensa que la principal fuente de ruido en estos procesos intracelulares son las fluctuaciones estadísticas en las concentraciones de ARNm y de proteinas reguladoras, las cuales afectan de manera directa la expresión de diferentes genes de acuerdo al caso.

El objetivo de este capítulo es introducir, basándonos en el capítulo *Noise in gene regulatory networks* del libro 'Complex Systems Science In Biomedicine' [20], la poderosa técnica de modelamiento analítico que es el acercamiento a un sistema por medio de la ecuación maestra. Este enfoque puede ser usado tanto para calcular propiedades estadísticas del ruido de un sistema, como para hallar los momentos de la distribución del compuesto que se esté siguiendo en el sistema; con ellos se puede incluso hacer una estimación de dicha distribución.

En una sola célula las concentraciones de estos compuestos pueden variar bastante puesto que muchas de estas moléculas están presentes en bajos números. Por esto la ecuación maestra es tan útil, ya que describe cómo cambia la probabilidad de estar en cierto estado del sistema en el tiempo. Por ejemplo, si una molécula A produce una molécula B a una tasa k (en unidades de $[concentración * tiempo]^{-1}$) la ecuación se compone de los términos que decrecen la probabilidad pues representan una cambio de la configuración de estudio, como la transición $[a, b] \to [a, b + 1]$; y términos que aumentan la probabilidad pues representan un cambio de una configuración diferente hacia la estudiada, como la transición $[a, b - 1] \to [a, b]$. Luego se tiene para dicha reacción:

$$\frac{d}{dt}P(a, b, t) = -kaP(a, b, t) + kaP(a, b - 1, t) \tag{3}$$

Donde a y b son el número de moléculas de A y de B respectivamente. Si ahora quisiéramos tener en cuenta la degradación de la molécula B, a una tasa γ medida también en $[concentración * tiempo]^{-1}$, deberiamos agregar los términos que suman y restan probabilidad respecto a este proceso, es decir

$$\gamma(b + 1)P(a, b + 1, t) \qquad \gamma bP(a, b, t) \tag{4}$$

Luego la nueva ecuación maestra sería:

$$\frac{d}{dt}P(a, b, t) = -kaP(a, b, t) + kaP(a, b-1, t) + \gamma(b+1)P(a, b+1, t) - \gamma bP(a, b, t) \tag{5}$$

Ahora viene lo interesante: para poder sacar los momentos de la distribución de A o de B se hace conveniente definir la siguiente función:

$$F(z_1, z_2, t) = \sum_{a,b} z_1^a z_2^b P(a, b, t) \tag{6}$$

La suma se hace sobre todo el dominio de A y de B. A esta función se le llama la función generadora de momentos por la siguiente propiedad:

$$\frac{\partial^q}{\partial z_n^q} F(z_1, z_2, ..., z_m, t)|_{z_n=1} = \left\langle \frac{a_n!}{(a_n - q)!} \right\rangle \tag{7}$$

En donde he puesto m variables a_n nuevas para hacer énfasis en que se puede definir una nueva por cada uno de los compuestos presentes en el sistema, pues estos describen una configuración específica (son relevantes).

Luego podemos multiplicar la ecuación maestra para el sistema en cuestión por $z_1^a z_2^b$ y sumar sobre a y b para obtener una ecuación para la función generadora de momentos, y por ende para los momentos.

$$\dot{F}(z_1, z_2, t) = k z_1 (z_2 - 1) \frac{\partial}{\partial z_1} F(z_1, z_2, t) - \gamma (z_2 - 1) \frac{\partial}{\partial z_2} F(z_1, z_2, t) \tag{8}$$

A dicha ecuación se llega después de hacer algunos cambios de variable y jugar con los límites de las sumas; pero ¡no os preocupéis! Estos procedimientos serán explorados a mucho detalle en las siguientes secciones.

3. Primer acercamiento al sistema

Cabe resaltar que el enfoque de este trabajo es estudiar un proceso de expresión genética en abstracto, por lo que como primer acercamiento veremos el sistema de producción constitutiva (en otras palabras, durante cualquier

condición fisiológica) del compuesto g. En este proceso se toma en cuenta la descomposición y creación de moléculas de proteina g.

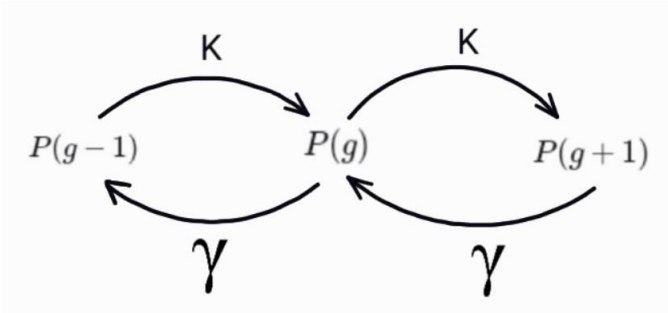

Figura 4: Esquema de la ecuación maestra del sistema de producción constitutiva y degradación del compuesto g.

Para ver como varía g debemos hacerlo por medio de la ecuación maestra. La gracia de tomar este sistema es ilustrar como resolverlo para la situación fuera del estado estable. Veremos que el compuesto s varía en el tiempo bajo una distribución de probabilidad, la cual se puede hallar por medio de la manipulación de la ecuación maestra. Por esta razón se vuelve importante ver la evolución temporal de esta distribución. La ecuación maestra queda de la siguiente manera:

$$\frac{d}{dt}P(g) = kP(g-1) - \gamma g P(g) - kP(g) + \gamma(g+1)P(g+1) \qquad (9)$$

3.1. Modificación de la ecuación maestra

Para intentar expresar la ecuación en términos de la ecuación generadora de momentos $\sum_{g=g_{min}}^{g_{max}} z^g P(g) = F$, multiplicamos por $\sum_{g=g_{min}}^{g_{max}} z^g$. A partir de lo anterior se puede modificar la ecuación de término a término de esta manera:

- $\sum_{g=g_{min}}^{g_{max}} z^g kP(g-1) = \sum_{g=g_{min}+1}^{g_{max}} z^g kP(g-1) = \sum_{n=g_{min}}^{g_{max}-1} z^{n+1} kP(n) = zk \sum_{n=g_{min}}^{g_{max}-1} z^g P(n)$
  $\approx zk \sum_{n=g_{min}}^{g_{max}} z^n P(n) = zkF$

El último paso se da gracias a que se asume que $P(g_{max}) \approx 0$.

- $\sum_{g=g_{min}}^{g_{max}} z^g \gamma g P(g) = \sum_{g=g_{min}}^{g_{max}} z(\frac{d}{dz} z^g)\gamma P(g) = \gamma z \frac{d}{dz} F$
- $\sum_{g=g_{min}}^{g_{max}} z^g \gamma (g+1) P(g+1) = \sum_{g=g_{min}+1}^{g_{max}+1} z^{n-1} \gamma n P(n) = \gamma \frac{d}{dz} \sum_{g=g_{min}}^{g_{max}} z^n P(n)$ $= \gamma \frac{d}{dz} F$

  Donde se tuvo en cuenta que $P(g_{max}+1) = 0$

De lo anterior se obtiene:

$$\dot{F} = zkF - \gamma z \frac{d}{dz} F - kF + \gamma \frac{d}{dz} F \tag{10}$$

$$\dot{F} = k(z-1) - \gamma(z-1)\frac{dF}{dz} \tag{11}$$

3.2. Aplicación de la función generadora de momentos

La acuación a la que se llegó resulta muy últil, pues al usar $F|_{z=1} = 1$ y $\frac{dF}{dz}|_{z=1} = \langle g \rangle$ se llega a la siguiente ecuación para el valor esperado del compuesto g:

$$\dot{\langle g \rangle} = k - \gamma \langle g \rangle \tag{12}$$

Ecuación cuya solución es:

$$\langle g \rangle = C_1 e^{-\gamma t} + \frac{k}{\gamma} \tag{13}$$

De donde podemos ver la tendencia al estado estable $\langle g \rangle = k/\gamma$ para tiempos

grandes. La importancia de esta ecuación no se limita a solo eso, también es posible sacar todos los demás momentos $\langle g^n \rangle$ de la distribución de probabilidad $P(g)$ en función del tiempo a partir de ella. Para ver esto primero es importante notar que

$$\frac{d^n}{dz^n}[(z-1)\frac{d^m}{dz^m}F] = n\frac{d^{n+m-1}}{dz^{n+m-1}}F + (z-1)\frac{d^{n+m}}{dz^{n+m}}F \tag{14}$$

$$\Rightarrow \frac{d^n}{dz^n}[(z-1)\frac{d^m}{dz^m}F]|_{z=1} = n\Big\langle \frac{g!}{(g-m-n+1)!}\Big\rangle \tag{15}$$

Luego, podemos construir una ecuación diferencial para todo $n \in \mathbb{N}$:

$$\frac{d^n}{dz^n}\dot{F} = nk\frac{d^{n-1}F}{dz^{n-1}} - n\gamma\frac{d^nF}{dz^n} + k(z-1)\frac{d^nF}{dz^n} - \gamma(z-1)\frac{d^{n+1}F}{dz^{n+1}} \tag{16}$$

$$\Rightarrow \Big\langle \frac{\dot{g}!}{(g-n)!}\Big\rangle = nk\Big\langle \frac{g!}{(g-n+1)!}\Big\rangle - n\gamma\Big\langle \frac{g!}{(g-n)!}\Big\rangle \tag{17}$$

Lo cual genera una relación de recurrencia para todos los momentos $\langle g^n \rangle$ con base en el primero. Lo importante de los momentos es que con ellos se puede construir la función caracteristica, y a partir de esta es posible calcular la función de probabilidad $P(g)$. Para este caso, todos los términos con dependencia temporal de los momentos serán exponenciales decadentes en el tiempo, dejando así una distribución de poisson (con parámetro k/γ) en la variable g para tiempos muy grandes.

4. Sistema más real

Como segundo acercamiento le agregamos al sistema anterior un término de *feedback* para lograr modelar la represión o activación del compuesto deacuerdo a su misma concentración. En particular esto nos acerca más al estudio de la acción del ambiente (s) sobre un sistema en general con esta forma.

Lo anterior no quiere decir que este tipo de sistema no se halla estudiado,

de hecho M. Thattai y Alexander van Oudenaarden en el 2002 estudiaron sistemas de este tipo [21]. Sus estudios se basaron en sistemas de cascadas de señales en redes reguladoras de genes, en donde las reacciones presentes en estas cascadas presentaban presentaban fluctuaciones aleatorias, dejando ruido en los diferentes niveles de señales de salida en la cascada. Sin embargo, en el planteamiento propuesto en ese estudio los canales aleatorios se modelan como ruido gaussiano blanco, y se toma la acción de un compuesto en el siguiente (bajo el orden de la cascada) como proporcional al compuesto anterior, aproximando de manera lineal los términos de represión o activación. Lo anterior provee una ventaja para el enfoque de Langevin tomado, sin embargo hace que se pierda exactitud. Para evitar esto en el presente estudio se propuso definir rangos de validez para los cuales aún así con las aproximaciones que se hagan el error se mantenga bajo.

Un término de feedback representa a la fracción de sitios activos que estan ocupados por un ligando en la proteina receptora; tiene la forma $\frac{\beta}{1+(\frac{r}{k})^h}$ para un sustrato r. Donde β es el número máximo de moléculas del sustrato que se convierten en producto por segundo, k es la concentración de sustrato a la cual la velocidad de reacción es la mitad de la velocidad máxima β, y h es la constante de Hill (medida de la cooperatividad). En un proceso de activación, cuanto más alto es el valor de h, mayor es el grado de cooperatividad. Número de sitios activos $> h$. Si $h = 1$ no hay cooperatividad; si $h > 1$, hay cooperatividad positiva, de lo contrario es cooperatividad negativa.

Agregando estos términos a la ecuación maestra

$$\begin{aligned}\frac{d}{dt}P(g) = \; & aP(g-1) - aP(g) + \frac{\beta}{1+(\frac{g-1}{k})^h}P(g-1) \\ & - \frac{\beta}{1+(\frac{g}{k})^h}P(g) - \gamma g P(g) + \gamma(g+1)P(g+1)\end{aligned} \tag{18}$$

4.1. Definiciones pertinentes

A continuación haremos algunas definiciones que facilitarán (y más aún, permitirán) el análisis del sistema. Primero está el cambio de variable alea-

toria de g a α y la definición de la nueva función generadora de momentos $\tilde{F}(z)$.

$$\alpha \equiv \left(\frac{g}{k}\right)^h \qquad y \qquad \tilde{F}(z) \equiv \sum_{\alpha} z^{\alpha} \tilde{P}(\alpha) = \sum_{g=g_{min}}^{g_{max}} z^{(\frac{g}{k})^h} P(g) \tag{19}$$

Lo anterior es cierto puesto que son distribuciones de probabilidad discretas y a su vez la variable g presenta una transformación uno a uno $g \to \left(\frac{g}{k}\right)^h$, donde $dom[P(g)] \subseteq \mathbb{N}$. La variable g puede ser renormalizada para estudios más específicos dependiendo del sistema tratado ($g \to cg$, con c constante).

Este cambio de variable nos permite llegar a la distribución de probabilidad de g deseada en ultima instancia. Se puede lograr hallando los momentos para la distribución de probabilidad de la nueva variable α y, a partir de la información que estos nos otorgan de la distribución de α (ya sea reconstruyendo dicha distribución a partir de estos o haciendo un análisis momento a momento), reversar el cambio de variable para obtener la distribución de probabilidad de la concentración g de proteinas.

4.2. Términos de feedback

Transformaremos entonces la ecuación maestra a una nueva ecuación para la nueva función generadora de momentos $\tilde{F}(z)$. Esta función cumple con las mismas propiedades que la función generadora de momentos usada en la sección anterior, salvo que esta genera los momentos $\langle \alpha^n \rangle$ de la distribución de probabilidad $\tilde{P}(\alpha)$. Multiplicando a ambos lados por z^{α} y sumando sobre todo el dominio de g (o α) tenemos:

$$
\begin{aligned}
\bullet \sum_{g=g_{min}}^{g_{max}} \frac{\beta}{1+(\frac{g}{k})^h} z^{(\frac{g}{k})^h} P(g) &= \sum_{\alpha} \frac{\beta}{1+\alpha} z^{\alpha} \tilde{P}(\alpha) \\
&= \sum_{\alpha} \sum_{n=0}^{\infty} (-1)^n 2^{-1-n} (\alpha-1)^n \beta z^{\alpha} \tilde{P}(\alpha) \\
&\approx \sum_{\alpha} \frac{\beta}{2} z^{\alpha} \tilde{P}(\alpha) \Big[1 - \frac{1}{2}(\alpha-1) + \frac{1}{4}(\alpha^2 - 2\alpha + 1) \Big] \\
&= \sum_{\alpha} \frac{\beta}{2} z^{\alpha} \tilde{P}(\alpha) \Big[\frac{7}{4} - \alpha + \frac{\alpha^2}{4} \Big] \\
&= \frac{\beta}{2} \Big[\frac{7}{4} - z\frac{\partial}{\partial z} + \Big(\frac{z}{2} \frac{\partial}{\partial z} \Big)^2 \Big] \sum_{\alpha} z^{\alpha} \tilde{P}(\alpha) \\
&= \frac{\beta}{2} \Big[\frac{7}{4} - z\frac{\partial}{\partial z} + \Big(\frac{z}{2} \frac{\partial}{\partial z} \Big)^2 \Big] \tilde{F}(z)
\end{aligned}
\tag{20}
$$

Para esto se usó la expansión $\frac{\beta}{1+\alpha} = \sum_{n=0}^{\infty} (-1)^n 2^{-1-n} (\alpha-1)^n$ la cual es válida para $|1-\alpha| < 2$, lo que acota un rango para α de $-1 < \alpha < 3$. A partir de lo anterior es posible especificar un rango de validez para la contante de Hill, dependiendo del rango que se quiera estudiar a detalle de la variable g con respecto a k (constante de disociación). Por ejemplo, si se toma como máximo valor $g_{max} = 2k$ luego $\left(\frac{g}{k}\right)^h < 2^h = \alpha_{max} < 3 \Rightarrow h < \frac{ln(3)}{(2)} \approx 1,585$. En general se tiene:

$$
h < \frac{ln3}{ln(\frac{g_{max}}{k})} \tag{21}
$$

$$
\begin{aligned}
\bullet \sum_{g=g_{min}}^{g_{max}} \frac{\beta}{1+(\frac{g-1}{k})^h} z^{(\frac{g}{k})^h} P(g-1) &= \sum_{n=g_{min}-1}^{g_{max}-1} z^{(\frac{n+1}{k})^h} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&= \sum_{n=g_{min}}^{g_{max}-1} z^{(\frac{n+1}{k})^h} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&\approx \sum_{n=g_{min}}^{g_{max}-1} z^{\left[(\frac{n}{k})^h + \frac{h}{k^h} n^{h-1}\right]} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&= \sum_{n=g_{min}}^{g_{max}-1} z^{\frac{h}{k}(\frac{n}{k})^{h-1}} z^{(\frac{n}{k})^h} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&\approx \sum_{n=g_{min}}^{g_{max}-1} z^{\frac{h}{k}} z^{(\frac{n}{k})^h} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&\approx z^{\frac{h}{k}} \sum_{n=g_{min}}^{g_{max}} z^{(\frac{n}{k})^h} \frac{\beta}{1+(\frac{n}{k})^h} P(n) \\
&= z^{\frac{h}{k}} \sum_{\alpha} \frac{\beta}{1+\alpha} z^{\alpha} \tilde{P}(\alpha) \\
&\approx z^{\frac{h}{k}} \frac{\beta}{2} \Big[\frac{7}{4} - z\frac{\partial}{\partial z} + \Big(\frac{z}{2} \frac{\partial}{\partial z} \Big)^2 \Big] \tilde{F}(z)
\end{aligned}
\tag{22}
$$

En donde se tomó en cuenta que $(\frac{g+1}{k})^h = (\frac{g}{k})^h [1+\frac{1}{g}]^h \approx (\frac{n}{k})^h + \frac{h}{k}(\frac{n}{k})^{h-1}$, puesto que para toda $g \in dom[P(g)]$ se tiene que $1/g \ll 1$. También se hizo la aproximación $(\frac{g}{k})^{h-1} \approx 1$ para poder sacar el factor $z^{\frac{h}{k}}$ de la suma, lo que permite convertir todo el término en el estudiado anteriormente. La anterior aproximación acota aún más el rango de validez para h, pues ahora h solo puede tomar valores positivos cercanos a 1. Es posible calcular el error para esta aproximación, este dependerá del dominio de g en comparación con k, o del verdadero valor de h si se tiene. El último paso, en donde se amplía el rango de la suma, se hace gracias a que $P(g_{max}) \ll 1$.

Un aspecto importante a notar de las aproximaciones anteriores es que condicionan a que este estudio solo logre describir de manera coherente el comportamiento de una función de represión del compuesto g.

4.3. Términos de degradación y producción basal

En cuanto a los términos de degradación:

- $$\sum_{g=g_{min}}^{g_{max}} z^{(\frac{g}{k})^h}\gamma g P(g)$$
$$= \sum_{\alpha} \gamma k \alpha^{\frac{1}{h}} z^{\alpha} \tilde{P}(\alpha)$$
$$\approx \sum_{\alpha} \gamma k z^{\alpha}\Big[1 + \frac{1}{h}(\alpha-1) + \frac{1}{h}\Big(\frac{1}{h}-1\Big)\frac{(\alpha-1)^2}{2} + \frac{1}{h}\Big(\frac{1}{h}-1\Big)\Big(\frac{1}{h}-2\Big)\frac{(\alpha-1)^3}{6}\Big]\tilde{P}(\alpha)$$
$$= \sum_{\alpha} \gamma k z^{\alpha}\Big[c_0 + c_1\alpha + c_2\frac{\alpha^2}{2} + c_3\frac{\alpha^3}{6}\Big]\tilde{P}(\alpha)$$
$$= k\gamma\Big[c_0 + c_1\Big(z\frac{\partial}{\partial z}\Big) + \frac{c_2}{2}\Big(z\frac{\partial}{\partial z}\Big)^2 + \frac{c_3}{6}\Big(z\frac{\partial}{\partial z}\Big)^3\Big]\sum_{\alpha} z^{\alpha}\tilde{P}(\alpha)$$
$$= k\gamma\Big[c_0 + c_1\Big(z\frac{\partial}{\partial z}\Big) + \frac{c_2}{2}\Big(z\frac{\partial}{\partial z}\Big)^2 + \frac{c_3}{6}\Big(z\frac{\partial}{\partial z}\Big)^3\Big]\tilde{F}(z) \tag{23}$$

$$
\begin{aligned}
&\bullet \sum_{g=g_{min}}^{g_{max}} z^{(\frac{g}{k})^h}\gamma(g+1)P(g+1) \\
&= \sum_{n=g_{min}+1}^{g_{max}+1} z^{(\frac{n-1}{k})^h}\gamma n P(n) \\
&\approx \sum_{n=g_{min}+1}^{g_{max}} z^{(\frac{n}{k})^h} z^{-\frac{h}{k}(\frac{n}{k})^{h-1}}\gamma n P(n) \\
&\approx \sum_{n=g_{min}+1}^{g_{max}} z^{-\frac{h}{k}} z^{(\frac{n}{k})^h}\gamma n P(n) \\
&= z^{-\frac{h}{k}} \sum_{\alpha} z^{\alpha}\gamma k \alpha^{\frac{1}{h}}\tilde{P}(\alpha) \\
&\approx z^{-\frac{h}{k}} k\gamma\Big[c_0 + c_1\Big(z\frac{\partial}{\partial z}\Big) + \frac{c_2}{2}\Big(z\frac{\partial}{\partial z}\Big)^2 + \frac{c_3}{6}\Big(z\frac{\partial}{\partial z}\Big)^3\Big]\tilde{F}(z)
\end{aligned}
\tag{24}
$$

En donde se usaron aproximaciones similares a las del segundo término de represión estudiado, con la diferencia que aquí se toma como pequeño $P(g_{min})$. Los valores de c_0, c_1, c_2 y c_3 se encuentran determinados por la expansión de $\alpha^{\frac{1}{h}}$ centrada en 1 hasta tercer orden:

$$
\begin{aligned}
c_0 &= 1 - \frac{11}{6h} - \frac{1}{6h^2} & \qquad c_1 &= \frac{3}{h} - \frac{5}{2h^2} + \frac{1}{2h^3} \\
c_2 &= -\frac{3}{h} + \frac{4}{h^2} - \frac{1}{h^3} & \qquad c_3 &= \frac{2}{h} - \frac{3}{h^2} + \frac{1}{h^3}
\end{aligned}
\tag{25}
$$

Por último, los terminos de producción basal:

$$
\bullet \sum_{g=g_{min}}^{g_{max}} z^{(\frac{g}{k})^h} a P(g) = \sum_{\alpha} z^{\alpha} a\tilde{P}(\alpha) = a\tilde{F}(z) \tag{26}
$$

$$
\begin{aligned}
\bullet \sum_{g=g_{min}}^{g_{max}} z^{(\frac{g}{k})^h} aP(g-1) &= \sum_{n=g_{min}-1}^{g_{max}-1} z^{(\frac{n+1}{k})^h} aP(n) \\
&\approx \sum_{g=g_{min}}^{g_{max}-1} z^{(\frac{n}{k})^h} z^{\frac{h}{k}\frac{n}{k}^{h-1}} aP(n) \\
&\approx \sum_{g=g_{min}}^{g_{max}} z^{\frac{h}{k}} a z^{(\frac{n}{k})^h} P(n) \\
&= a z^{\frac{h}{k}} \sum_{\alpha} \tilde{P}(\alpha) = a z^{\frac{h}{k}} \tilde{F}(z)
\end{aligned}
\tag{27}
$$

4.4. Ecuación final

Teniendo todos los términos transformados, ahora la nueva ecuación queda así:

$$
\begin{aligned}
\frac{d}{dt}\tilde{F}(z) =& \frac{\beta}{2}(z^{\frac{h}{k}} - 1)\Big[\frac{7}{4} - \Big(z\frac{\partial}{\partial z}\Big) + \Big(\frac{z}{2}\frac{\partial}{\partial z}\Big)\Big]\tilde{F}(z) \\
&+ k\gamma(z^{-\frac{h}{k}} - 1)\Big[c_0 + c_1\Big(z\frac{\partial}{\partial z}\Big) + \frac{c_2}{2}\Big(z\frac{\partial}{\partial z}\Big)^2 + \frac{c_3}{6}\Big(z\frac{\partial}{\partial z}\Big)^3\Big]\tilde{F}(z) \\
&+ a(z^{\frac{h}{k}} - 1)\tilde{F}(z)
\end{aligned}
\tag{28}
$$

Cabe notar que al igual que la ecuación para la función generadora de momentos que fue hallada en la sección anterior, esta también se anula cuando evaluamos $z = 1$ antes de derivar. Al derivar y evaluar en $z = 1$ se obtiene la siguiente expresión:

$$
\begin{aligned}
\frac{d}{dt}\langle\alpha\rangle =& \frac{\beta}{2}\frac{h}{k}\Big[\frac{7}{4} - \langle\alpha\rangle + \frac{\langle\alpha^2\rangle}{4}\Big] + a\frac{h}{k} \\
&- k\gamma\frac{h}{k}\Big[c_0 + c_1\langle\alpha\rangle + \frac{c_2}{2}\langle\alpha^2\rangle + \frac{c_3}{6}\langle\alpha^3\rangle\Big]
\end{aligned}
\tag{29}
$$

Dado que en esta expresión tampoco se tiene en cuenta el ambiente externo cambiando, es posible deducir que las contribuciones a los momentos dependientes del tiempo irán decreciendo hasta que sean casí cero. Por esta misma razón es posible seguir con el análisis planteado en estado estable ($\frac{d}{dt}\langle\alpha\rangle = 0$). Tomando la expresión en estado estable, esta nos muestra una relación entre los tres primeros momentos de la distribución $\tilde{P}(\alpha)$.

4.5. Conexión con teoría de la información

En esta área hay muchas cosas que falta por experimentar y modelar. Una de ellas, y que llamó mucho mi atención, fue la búsqueda por límites para el nivel al que se puede aumentar el fitness de una población de organismos unicelulares. Pues este problema encarna el buscar qué tanto se puede maximizar una figura de mérito representativa de una población de seres vivos de acuerdo a su entorno; cosa que sería interesante probar que es posible hacerlo, y quién sabe, hasta poder plantearlo en humanos (no puede negar lector que la idea suena divertida). En este trabajo se pretendió dar uno de los pasos que nos puede acercar a una conclusión sobre este estudio dado que el modelo de expresión genética donde un compuesto interactúa consigo mismo es similar al modelo de un sistema de expresión genética en donde un compuesto del ambiente s reprime o coopera con una proteina del sistema g.

A partir de dicho modelo es posible hallar la distribución de pobabilidad de g dada la concentración del compuesto s, $P(g|s)$, para construir la información mutua entre ambiente y sistema dada la fórmula vista en la sección de formalización de conceptos. Lo único faltante para construir la información mutua sería la distribución de probabilidad del compuesto s del ambiente, $P(s)$. Esta distribución es algo complicado de hallar, incluso acercarse a ella de manera experimental es muy dificil puesto que requeriría medidas muy exactas y prolongadas del medio en el que vivan generaciones y generaciones de la población de organismos unicelulares que se quieran modelar. Sin embargo se pueden configurar diferentes ambientes en el laboratorio para los organismos unicelulares, incluso que estos cambien de manera aleatoria bajo una distribución, por lo que es factible asumir una distribución $P(s)$ con el fin de contruir la expresión de la información mutua.

Una vez se construya tal expresión lo único que se debe hacer es minimizarla en torno a las variables que pueden cambiar los procesos evolutivos (por ejemplo constantes de producción basal del compuesto g), imponiendo la condición de un fitness medio $< f >$ para hallar la función $I_{min}(\langle f \rangle)$. Invirtiendo esta función se tendría la función $\langle f \rangle_{máx}(I)$ que nos dice cuál sería el fitness máximo de la población unicelular modelada dada la cantidad de información que esta extraiga del ambiente.

5. Conclusiones

El desarrollo del método bajo el orden de exactitud que se desee (por medio de las aproximaciones en las expansiones en series) lo convierte en una herramienta innovadora y eficaz para el estudio analítico de sistemas de expresión genética. Una de las caracteristicas importantes a notar es que no solo se puede minimizar el error por medio del número de términos que se escojan en las expansiones en series, sino que también se pueden definir de manera exacta rangos de validez para las aproximaciones que se hacen.

Es importante notar que en efecto el método usado es muy útil para hallar los momentos de la distribución $\tilde{P}(\alpha)$, lo cual nos lleva en últimas a los momentos de la distribución $P(g)$. A partir de estos es posible construir la distribución de probabilidad por medio de varias maneras: una sería un cálculo de minimización de la entropía con los momentos como restricciones; y la otra una aproximación a la distribución de probabilidad usando expansiones de Edgeworth, que basicamente es una aproximación a la distribución en término de sus cumulantes.

Lo único necesario para tener una aproximación tan buena como queramos de la distribución de probabilidad en este caso son los valores de $\langle \alpha \rangle$ y $\langle \alpha^2 \rangle$. Con estos valores dada la ecuación para $\frac{d}{dt}\tilde{F}(z)$ se tiene una relación de recurrencia para todos los demás momentos.

Vale la pena mirar el problema intentando generalizar para valores de la constante de Hill (h) negativos, puesto que lastimosamente una de las aproximaciones más importantes llevó al requisito que h debía ser parecido a 1.

Referencias

[1] CE Shannon. A mathematical theory of communication. *Bell Sys Tech J*, 27:379–423 and 623–656, 1948.

[2] CE Shannon. Communication in the presence of noise. *Proc IRE*, 37: 10–21, 1949.

[3] George A Miller. What is information measurement? *American Psychologist*, 8(1):3–11, Jan 1953.

[4] F Attneave. Some informational aspects of visual perception. *Psych Rev*, 61:183–193, 1954.

[5] HB Barlow. Possible principles underlying the transformation of sensory messages. in sensory communication. *W Rosenblith(MIT Press, Cambridge,)*, pages 217–234, 1961.

[6] F Rieke W Bialek, M DeWeese and D Warland. Bits and brains: Information flow in the nervous system. *Physica A*, 200:581–593, 1993.

[7] M DeWeese and W Bialek. Information flow in sensory neurons. *Il Nuovo Cimento*, 17 D:733–742, 1995.

[8] R de Ruyter van Steveninck F Rieke, D Warland and W Bialek. Spikes: Exploring the neural code. *MIT Press,Cambridge*, 1997.

[9] RR de Ruyter van Steveninck SP Strong, R Koberle and W Bialek. Entropy and information in neural spike trains. *Phys Rev Lett*, 80:197–200, 1998.

[10] W Bialek SP Strong, RR de Ruyter van Steveninck and R Koberle. On the application of information theory to neural spike trains. *Pacific Symposium on Biocomputing '98, RB Altman, AK Dunker, L Hunter and TE Klein (World Scientific, Singapore, 1998).*, pages 621–632, 1998.

[11] I Nemenman E Ziv and C Wiggins. Optimal signal processing in small stochastic biochemical networks. *PLoS One*, 2, 2007. doi: arXiv:q–bio.MN/061204.

[12] CG Callan Jr G Tkacik and W Bialek. Information flow and optimization in transcriptional regulation. *Proc Nat Acad Sci(USA)*, 105: 12265–12270, 2008. doi: arXiv:0705.0313 [q–bio.MN].

[13] CG Callan Jr G Tkacik and W Bialek. Information capacity of genetic regulatory elements. *Phys Rev E*, 78, 2008. doi: arXiv:0709.4209 [q–bio.MN] (2007).

[14] W Bialek and S Setayeshgar. Cooperativity, sensitivity and noise in biochemical signaling. *Phys Rev Lett*, 100, 2008. doi: q–bio.MN/0601001.

[15] AM Walczak G Tkacik and W Bialek. Optimizing information flow in small genetic networks. iii. a self-interacting gene. *Phys Rev E*, 85, 2012. doi: arXiv.org:1112.5026 [q–bio.MN].

[16] CG Callan Jr G Tkacik and W Bialek. Information capacity of genetic regulatory elements. *Phys Rev E*, 78, 2008. doi: arXiv:0709.4209 [q–bio.MN] (2007).

[17] N Tishby SF Taylor and W Bialek. Information and fitness. *Proc Nat Acad Sci(USA)*, 2007. doi: arXiv:0712.4382 [q–bio.PE].

[18] Edu Kussell and Stanislas Leibler. Phenotypic diversity, populat growth, and information in fluctuating environments. *Science*, 309, 2005. doi: DOI: 10.1126/science.1114383.

[19] E. Dekel and U. Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436, 2005. doi: doi:10.1038/nature03842.

[20] Alexander van Oudenaarden Juan M. Pedraza. *Complex Systems Science In Biomedicine: Noise in gene regulatory networks*. Springer, 2006.

[21] Mukund Thattai and Alexander van Oudenaarden. Attenuation of noise in ultrasensitive signaling cascades. *Biophysical Journal*, 82, 2002.