ScienceDirect
Disponible en www.sciencedirect.com
Revista Iberoamericana de Automática e Informática industrial 12 (2015) 241–252
www.elsevier.es/RIAI

Modelado de sistemas bioquímicos: de la Ley de Acción de Masas a la Aproximación Lineal del Ruido

Jesús Picó[a,*], Alejandro Vignoni[b], Enric Picó-Marco[a], Yadira Boada[a]

[a] *Instituto Universitario de Automática e Informática Industrial, Universitat Politècnica de València, C/ Camí de Vera, nº 14, 46022, València, España.*
[b] *Centro de Biología de Sistemas, Instituto Max Planck de Biología Celular Molecular y Genética, C/Pfotenhauerstr. nº 108, 01307, Dresden, Alemania*

Resumen

Durante la ultima década hemos vivido una creciente aplicación de técnicas propias de las ingenierías a la biología. Áreas como la Biología de Sistemas o, más recientemente, la Biología Sintética, reciben una atención cada vez mayor por parte de los ingenieros. En particular, el modelado en estos ámbitos permite la generación de nuevas hipótesis contrastables experimentalmente, y de nuevas formas de intervención biológica, así como explicaciones más o menos mecanicistas de los resultados experimentales. Una aproximación basada en modelo requiere considerar la dinámica de las reacciones bioquímicas y su regulación. En la primera parte de este tutorial se introducen el modelado determinista y reducción de modelos de la clase de reacciones bioquímicas propias de la biología molecular celular.

El ruido juega un papel crucial en la dinámica de los circuitos biológicos. En el área de control automático hay una larga tradición de modelado mediante ecuaciones diferenciales estocásticas lineales, bajo la hipótesis simplificadora de asumir que el ruido tiene una magnitud independiente de la del estado. Esta hipótesis no es válida en los circuitos biológicos. En la segunda parte del tutorial se describen los métodos de modelado estocástico más usados en biología molecular, con especial atención a denominada aproximación lineal del ruido.

Palabras Clave:
Sistemas estocásticos, Ecuaciones diferenciales, Modelado de sistemas continuos, Reducción de modelos, Simulación de sistemas, Ruido, Sistemas biológicos, biotecnológicos y bioprocesos.

1. Introducción

Durante la ultima década hemos vivido una creciente aplicación de técnicas propias de las ingenierías a la biología. Áreas como la Biología de Sistemas o, más recientemente, la Biología Sintética, reciben una atención cada vez mayor por parte de los ingenieros. El diseño y producción clásicos en biotecnología, mediante el uso de organismos genéticamente modificados, ha seguido tradicionalmente una estrategia esencialmente de tipo prueba-y-error.

En los últimos años, esta aproximación estaba prácticamente agotada, llevando a un cuello de botella en la generación de nuevos productos biotecnológicos. Es en este punto cuando se ve clara la necesidad de aplicar modelado computacional, de optimizar computacionalmente en lugar de adivinar por prueba-y-error cada nueva ruta metabólica (Kwok, 2010). Este proceso de incorporación de herramientas de la ingeniería a la biología se ha intensificado tremendamente con el advenimiento de la Biología Sintética. Esta se define como la ingeniería de la biología: el (re)diseño y construcción deliberados de nuevos componentes, dispositivos y sistemas biológicos para realizar nuevas funciones con un propósito utilitario (De Lorenzo, 2014). Como disciplina ingenieril, la Biología Sintética persigue construir dispositivos que aún no existen, enfatizando los principios y metodologías de la ingeniería en el diseño, caracterización, y construcción de los mismos (Arpino et al., 2013; ERASynBio, 2014). En este contexto, las herramientas habilitadoras esenciales que subyacen a esta nueva ingeniería biológica son las propias de la biología molecular moderna –secuenciación, tecnologías -ómicas, síntesis y ensamblado de ADN– pero también la ingeniería de sistemas, y el modelado y diseño computacionales.

En particular, el modelado permite la generación de nuevas hipótesis contrastables experimentalmente, y de nuevas formas de intervención biológica, así como explicaciones más o menos mecanicistas de los resultados experimentales. Los modelos dinámicos son especialmente importantes para una aproxi-

*Autor en correspondencia.
Correos electrónicos: jpico@ai2.upv.es (Jesús Picó), vignoni@mpi-cbg.de (Alejandro Vignoni), enpimar@isa.upv.es (Enric Picó-Marco), yaboa@upv.es (Yadira Boada)
URL: http://sb2cl.ai2.upv.es (Jesús Picó)

http://dx.doi.org/10.1016/j.riai.2015.06.001

mación basada en modelo, pues permiten explicar y predecir el comportamiento que emerge de la evolución temporal de las concentraciones de los componentes celulares. Estos modelos requieren considerar la dinámica de las reacciones bioquímicas y su regulación. No en vano, el modelado cinético determinista de pequeños circuitos biológicos tiene una larga tradición (Chen et al., 2010; Villaverde and Banga, 2014).

Especial interés tiene el modelado en situaciones en las que el ruido juega un papel relevante. La necesidad de modelar el ruido se presenta en multitud de aplicaciones prácticas en las que un modelo determinista no es suficiente para capturar el comportamiento dinámico de los sistemas implicados con suficiente aproximación. En el ámbito de la Automática, normalmente bajo la asumción de que el ruido tiene una magnitud independiente de la del estado, hay una larga tradición de modelado de sistemas estocásticos mediante ecuaciones diferenciales estocásticas lineales (Aström, 2006). Sin embargo, en muchas aplicaciones, esta asumción simplificadora no es válida: la magnitud del ruido depende de la del estado. Es el caso de los modelos de reacciones bioquímicas usados en biología de sistemas y biología sintética.

En este tutorial se introducen este tipo de sistemas dinámicos y su modelado, partiendo del modelado determinista, para seguir con el modelado estocástico, poniendo especial énfasis en la denominada *aproximación lineal del ruido*. Esta última metodología permite *desacoplar* la dinámica de la evolución media del estado del sistema y la de la varianza del ruido que le afecta, lo que la convierte en una técnica de análisis muy útil no sólo en sistemas biológicos. Las distintas metodologías se ejemplifican vía un caso sencillo que servirá de hilo conductor: la transcripción no regulada de un gen.

2. Transcripción constitutiva de un gen

A lo largo de las siguientes secciones usaremos como ejemplo conductor el proceso de transcripción constitutiva, i.e. no regulada, de un gen. Se trata de un caso muy sencillo, pero que permite ver bien todas las metodologías que describiremos. El conocido como Dogma Central de la biología celular molecular (véase la figura 1) establece que para sintetizar una proteína (en el argot, *expresarla*), primero se debe *transcribir* a una molécula intermedia, el ARN mensajero, su secuencia codificadora en el gen correspondiente a la proteína (Alberts et al., 2009).

El proceso de transcripción lo lleva a cabo un complejo enzimático denominado ARN-polimerasa (RNAp). Esta se enlaza a una secuencia del ADN, llamada *promotor*, situada justo antes de la secuencia codificadora de la proteína. Una vez enlazada, la RNAp se desliza a lo largo de la secuencia del gen en el ADN, sintetizando la molecula de ARN mensajero (mRNA), hasta que se encuentra una secuencia de terminación en el gen. Por otro lado, la molécula de mRNA se degrada tanto por procesos activos como pasivos de degradación. Muchos de los valores para los parámetros involucrados en estas reacciones pueden encontrarse en bases de datos públicas como BioNumbers (Milo et al., 2010), o CyberCell (Sundararaj et al., 2004). Como muestra, para hacerse una idea de las escalas temporales de las que se está hablando, la tasa de transcripción para la

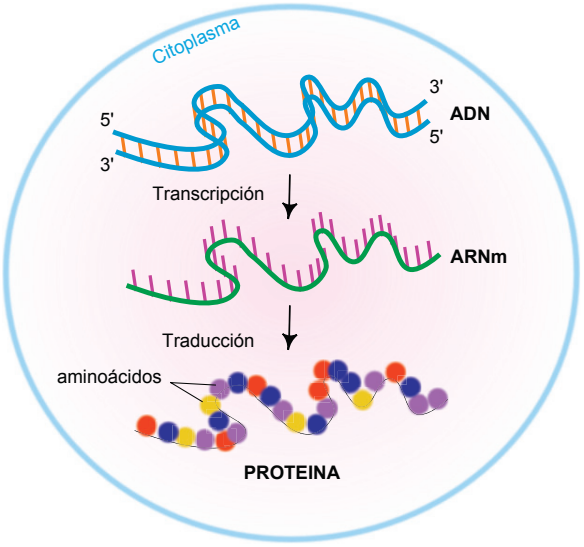

Figura 1: Dogma Central de la Biología, de ADN, a ARN mensajero, a Proteína.

bacteria *E. coli* está en el orden de [3, 10] moléculas de mRNA por minuto. La tasa de degradación, por otro lado, está en el intervalo [0,125; 1] min^{-1}.

En resumen, el proceso de transcripción esencialmente implica tres subetapas: (i) el enlace reversible de la RNAp al promotor del gen en el ADN, (ii) la elongación irreversible a lo largo del ADN para crear la copia de mRNA y (iii) la degradación (irreversible) del mRNA. Estas pueden expresarse como reacciones entre las especies bioquímicas involucradas:

$$\begin{aligned} \text{RNAp} + \text{DNA} &\underset{v_{-1}}{\overset{v_1}{\rightleftharpoons}} \text{RNAp} \cdot \text{DNA} \\ \text{RNAp} \cdot \text{DNA} &\xrightarrow{v_2} \text{RNAp} + \text{DNA} + \text{mRNA} \\ \text{mRNA} &\xrightarrow{v_d} \emptyset \end{aligned} \tag{1}$$

donde DNA representa el gen, y $RNAp \cdot DNA$ el complejo resultante del enlace entre la ARN polymerasa y el promotor del gen. Los símbolos v_1, v_{-1}, v_2, and v_d denotan las velocidades de reacción. Se asume que cuando la reacción de elongación finaliza, la ARN polimerasa se desprende del ADN, y el promotor queda libre para enlazarse de nuevo.

Este tipo de transcripción, en la cual el promotor sólo es activado por la ARN polimerasa se denomina constitutiva, o no regulada. En la mayor parte de los genes, el enlace de la ARN polimerasa al promotor puede verse afectado por la presencia de activadores o represores, de forma que la expresión de proteínas puede ser regulada en función del estado celular o del entorno.

3. Modelado determinista: la Ley de Acción de Masas

Todas las reacciones que tienen lugar en el interior celular son estocásticas por naturaleza. Por tanto, el modelado del conjunto de reacciones (1) debería ser formulado en términos de la probabilidad de que cada una de las reacciones tenga lugar. Los modelos estocásticos resultantes serán planteados en la sección

4 y posteriores. Los modelos deterministas, por otro lado, no tienen en cuenta la naturaleza probabilística de las reacciones. En su lugar, asumen que para cada especie química, la cantidad de moléculas transformadas por las reacciones depende sólo de la cantidad actual de moléculas, las velocidades de reacción, y la estequiometría de las reacciones. En este sentido, la Ley de Acción de Masas (LAM) es una formalismo ampliamente utilizado para expresar las velocidades de reacción de cada una de las reacciones en un sistema de reacciones como (1). El estado del sistema está constituido por el número de moléculas de cada especie o, alternativamente, por sus concentraciones.

La LAM establece que la velocidad de una reacción (bio)-química es proporcional al producto de las concentraciones de las especies reactantes elevadas a una potencia dada por la estequiometría de la reacción (Chellaboina et al., 2009). El factor de proporcionalidad es la tasa de reacción (velocidad específica de reacción). Esencialmente, la LAM asume que la velocidad de reacción es proporcional a la probabilidad de que los reactantes se encuentren (colisionen para reaccionar) y asocia esta probabilidad al producto de concentraciones. Si uno de los reactantes requeridos no está presente, la reacción no tendrá lugar. Por otro lado, la reacción procede a mayor velocidad si la concentración de los reactantes aumenta.

Consideremos las reacciones de transcripción (1). Siguiendo el formalismo introducido, se considera que las tres reacciones proceden con velocidades proporcionales al producto de las concentraciones de reactantes. Denotando $x_1 = [DNA]$, $x_2 = [RNAp]$, $x_3 = [RNAp{\cdot}DNA]$, y $x_4 = [mRNA]$, tendremos:

$$\begin{aligned} v_1 &= k_1 x_1 x_2 \\ v_{-1} &= k_{-1} x_3 \\ v_2 &= k_m x_3 \\ v_3 &= d_m x_4 \end{aligned} \tag{2}$$

donde las constantes de proporcionalidad k_1, k_{-1}, k_m, d_m son las velocidades de reacción específicas, o *tasas de reacción*. Con las velocidades (2) se pueden ahora establecer los balances de masa dinámicos para las cuatro especies:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} &= \begin{bmatrix} -1 & 1 & 1 & 0 \\ -1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \triangleq Av \\ &= \begin{bmatrix} -k_1 x_1 x_2 + k_{-1} x_3 + k_m x_3 \\ -k_1 x_1 x_2 + k_{-1} x_3 + k_m x_3 \\ k_1 x_1 x_2 - k_{-1} x_3 - k_m x_3 \\ k_m x_3 - d_m x_4 \end{bmatrix} \end{aligned} \tag{3}$$

donde A es la matriz de estequiometría.

3.1. Reducción del modelo determinista

El modelo dinámico (3) tiene la ventaja de ser un modelo dinámico positivo polinomial. Esto permite, en teoría, utilizar herramientas de análisis propias de los sistemas polinomiales, como el álgebra diferencial (Picó-Marco, 2013), o el Sum of Squares (SOS) para el análisis de estabilidad (Blanchini and Franco, 2011). En la práctica, sin embargo, el orden de los modelos obtenidos es muy elevado – para un proceso muy simple, como la transcripción constitutiva, se ha obtenido un modelo de cuarto orden. Un modelo de un circuito biológico básico puede fácilmente tener orden superior a unas cuantas decenas. Esto dificulta el uso de las mencionadas técnicas para sistemas polinomiales. Por consiguiente, en la práctica se prefieren modelos de orden reducido. Proceder a la reducción de modelos como (3) tiene además una serie de ventajas adicionales:

- las dificultades experimentales y computacionales para estimar parámetros en sistemas biológicos celulares se incrementan mucho con su número. Es más, en general es experimentalmente más factible estimar parámetros derivados resultantes de la agregación de otros más primitivos, ya que se suelen asociar a mecanismos biológicos de menor detalle, pero más fáciles de tratar experimentalmente.
- las dinámicas celulares son rígidas, existiendo una clara división de escalas temporales entre las distintas reacciones –típicamente de varios órdenes de magnitud– de la cual se puede sacar provecho.

En la práctica, conviene que el proceso de reducción proporcione un modelo en el que no se pierda relevancia biológica. En particular, suele evitarse que las especies del modelo reducido sean agregadas. Por otro lado, los parámetros agregados resultantes deben ser fácilmente asociables a mecanismos biológicos fáciles de estimar y/o modificar experimentalmente.

La reducción de modelos bioquímicos puede llevarse a cabo mediante aplicación de la denominada *Aproximación Quasi-Estacionaria* (QSSA por las siglas en inglés) de las reacciones químicas rápidas, y la obtención de relaciones algebraicas entre los estados del sistema.

En esencia, QSSA es un método de perturbaciones singulares (Hinch, 1991; Khalil, 2011; Kokotovic et al., 1986) que considera la separación de escalas temporales entre las diferentes dinámicas del sistema (Zagaris et al., 2004; Mélykúti et al., 2014). En concreto, se suele asumir que las dinámicas de las reacciones de enlace son muy rápidas en comparación con las de elongación y degradación. Además de herramienta para proceder a la reducción de modelos, el método es también muy útil a la hora de definir módulos y sus interacciones en circuitos biológicos complejos, con un formalismo semejante al uso de cuadripolos en electrotecnia o sistemas con puertos en análisis basado en pasividad, (Del Vecchio, 2013).

Las relaciones algebráicas entre estados se obtienen mediante la búsqueda de *invariantes* del sistema. En el caso de los sistemas de reacciones, algunas de ellas son combinación lineal de otras, lo cual implica que la combinación lineal de los estados correspondientes se mantendrán constantes en el tiempo. Estas combinaciones lineales de concentraciones[1] pueden ser entendidas como quasi-especies invariantes.

[1] Alternativamente puede expresarse el modelo como balances dinámicos sobre el número de moléculas de las especies.

Consideremos nuestro modelo de transcripción constitutiva (3). En este caso sencillo, los invariantes pueden obtenerse por simple inspección observando que $\dot{x}_1 + \dot{x}_3 = 0$, y $\dot{x}_2 + \dot{x}_3 = 0$, lo cual implica:

$$\begin{aligned} x_1 + x_3 &= c_n \\ x_2 + x_3 &= c_{\text{RNAp}} \end{aligned} \tag{4}$$

donde las constantes c_n, y c_{RNAp} son las concentraciones iniciales de $x_1 + x_3$, y $x_2 + x_3$ respectivamente. En sistemas de reacciones más grandes los invariantes se obtienen a partir del *kernel* de la matriz estequiométrica traspuesta A^T (Llaneras and Picó, 2008).

El primer invariante indica que el número de copias del gen, es decir la cantidad de ADN, se mantiene constante en el tiempo, y es igual a la suma de la cantidad de promotor (ADN) libre más el ocupado por la RNAp. La constante c_n es por tanto el número de copias del gen. Este es un parámetro biológico importante, que puede ser modificado experimentalmente.

El segundo invariante indica la conservación de la ARN polimerasa. La RNAp está bien libre (x_2), bien enlazada al ADN (x_3). Este invariante puede llevar a equívocos, y ha de interpretarse con cuidado. Implícitamente asume que el sistema bajo estudio está completamente aislado, lo cual no es cierto. La RNAp interviene en muchos procesos celulares simultáneamente. Es decir, la segunda ecuación del modelo (3) es una sobre-simplificación que asume que el único gen usando RNAp es el que estamos estudiando. El balance dinámico de la ARN polimerasa está, por tanto, incompleto. Pero no es factible introducir en el mismo todos los términos de todas las reacciones usando RNAp simultáneamente. Una buena opción para tratar esta situación de estado compartido por muchos subprocesos es considerar que la célula contiene suficiente RNAp libre en exceso como para *servir* a todos los genes activos transcribiendo simultáneamente. Bajo esta condición, la concentración de RNAp libre no cambiará apreciablemente en el tiempo y, por tanto, puede asumirse constante:

$$x_2 \approx c_{\text{RNApf}} \tag{5}$$

En la práctica, la RNAp libre puede cambiar apreciablemente en el tiempo si la célula está sometida a condiciones variantes. Pero este cambio sucede en una escala temporal muy lenta respecto a las que nos ocupan, por lo que c_{RNApf} puede ser considerado un parámetro lentamente variante en el tiempo.

Por otro lado, la aplicación del método QSSA nos permitirá eliminar los estados cuyas dinámicas son muy rápidas respecto a las de los estados restantes. La idea bajo el método QSSA es que las reacciones muy rápidas alcanzan estado estacionario muy rápidamente, por lo que podemos despreciar su dinámica y asumir que se encuentran en régimen permanente. Esto convertirá la ecuación diferencial correspondiente en una algebraica. Como ya indicamos, la formulación rigurosa del método, y la comprobación de las condiciones técnicas necesarias, pasa por el uso de método de perturbaciones singulares.

En nuestro ejemplo de transcripción se puede asumir con seguridad que las reacciones de enlace/desenlace de la ARN polimerasa al promotor del gen son mucho más rápidas que las de elongación y degradación. Esto se refleja en los valores de las correspondientes tasas de reacción. Recordando el modelo (3), podemos definir el parámetro de perturbación:

$$\epsilon = \frac{1}{k_1} \tag{6}$$

Pre multiplicando por ϵ en ambos lados de la primera ecuación del modelo se tiene:

$$\epsilon \dot{x}_1 = -x_1 x_2 + \frac{k_{-1} + k_m}{k_1} x_3 \tag{7}$$

Asumiendo k_1 suficientemente grande, ϵ será muy pequeño. En el límite, ϵ tenderá a cero conforme k_1 aumenta. El término $\frac{k_{-1}+k_m}{k_1}$ no se anulará si numerador y denominador tienen ordenes de magnitud comparables. Este es el caso para las tasas de enlace y desenlace de la RNAp al promotor. Bajo esta condición, (7) puede aproximarse como:

$$0 = -x_1 x_2 + \frac{k_{-1} + k_m}{k_1} x_3 \tag{8}$$

Mediante esta relación algebraica y los invariantes obtenidos anteriormente, llegamos al modelo reducido:

$$\dot{x}_4 = k_m \frac{c_n}{1 + \frac{1}{c_{\text{RNAp}}} \frac{k_{-1}+k_m}{k_1}} - d_m x_4 \tag{9}$$

donde recordemos que $x_4 = [\text{mRNA}]$.

El modelo obtenido muestra que la síntesis de ARN mensajero aumenta si el promotor tiene mucha afinidad por la ARN polimerasa (k_1 grande), hasta un límite dado por la tasas de la reacción de elongación y el número de copias del gen en la célula. También se observa que si hay muchos procesos celulares en paralelo consumiendo RNAp libre (c_{RNAp} bajo) la síntesis de ARN mensajero disminuirá.

El modelo (9) suele expresarse agregando todos sus parámetros bajo el paraguas de un nuevo parámetro que refleja la tasa de transcripción efectiva, y que es fácil de medir y modificar experimentalmente:

$$k_{me} = \frac{k_m}{1 + \frac{1}{c_{\text{RNAp}}} \frac{k_{-1}+k_m}{k_1}} \tag{10}$$

obteniéndose de este modo la expresión más usada para modelar la transcripción genética constitutiva:

$$\dot{x}_4 = k_{me} c_n - d_m x_4 \tag{11}$$

El proceso de transcripción (11) puede asociarse a un sistema de reacciones reducido equivalente:

$$\begin{aligned} \text{DNA} &\xrightarrow{k_{me}} \text{DNA} + \text{mRNA} \\ \text{mRNA} &\xrightarrow{d_m} \emptyset \end{aligned} \tag{12}$$

4. Modelado estocástico: la Ecuación Química Maestra

Estrictamente hablando, el proceso de generación y degradación de mRNA es estocástico. La Ecuación Química Maestra

(CME) establece el balance de las distribuciones de densidad de probabilidad[2]. A partir de este balance, se puede obtener la distribución de probabilidad de sistemas con n copias de mRNA en el instante t para cada instante de tiempo. En la práctica la CME sólo se puede resolver analíticamente para casos lineales simples. Por lo tanto, se usan bien aproximaciones numéricas bien aproximaciones analíticas.

Para los cuatro componentes en el conjunto de reacciones (1) podemos escribir las ecuaciones de balance para las distribuciones de probabilidad, dando lugar a la correspondiente CME. A continuación lo haremos sólo para el mRNA usando el modelo simplificado (12). Consideremos de nuevo la red de reacciones simplificada para transcripción:

$$\begin{aligned} \mathrm{DNA} &\xrightarrow{r} \mathrm{DNA} + \mathrm{mRNA} \\ \mathrm{mRNA} &\xrightarrow{d} \emptyset \end{aligned} \tag{13}$$

Ahora, en lugar de las tasas de reacción deterministas, se hablará de tasas de reacción probabilísticas. Más tarde veremos la relación entre las mismas. Por el momento, consideremos que la probabilidad por unidad de tiempo de que una molécula de mRNA sea transcrita es r, y la de que una molécula de mRNA se degrade es d. Asumamos que el número de copias del gen $c_n = 1$. Si este no es el caso, se podría considerar simplemente el producto rc_n como la probabilidad de la tasa de transcripción o, para ser más exactos, c_n reacciones paralelas como (13).

Denotemos la probabilidad de tener n copias de mRNA en el instante t como $p(n,t)$, y establezcamos un balance para obtener la probabilidad de tener n copias en el instante $t + \delta t$. El razonamiento es como sigue: la probabilidad que buscamos es igual a la probabilidad de tener $n-1$ copias de mRNA en t, habiendo tenido lugar la reacción de transcripción durante el intervalo de tiempo δt, mas la probabilidad de tener $n+1$ copias de mRNA en t, habiendo tenido lugar la reacción de degradación de mRNA durante el intervalo de tiempo δt, más la probabilidad de que ya hubieran n copias de mRNA en t, no habiendo tenido lugar ninguna reacción durante el intervalo de tiempo δt. Nótese que se usa adición de probabilidades porque se asume que tanto el primer evento, como el segundo, como el tercero, pueden ocurrir, y el tiempo transcurrido δt se toma lo suficientemente pequeño de manera que los tres posibles eventos sean disjuntos. Si dos eventos son mutuamente exclusivos, la probabilidad de que ocurra bien uno bien el otro es la suma de las probabilidades de que ocurra cada uno.

La probabilidad de que una reacción específica tenga lugar en un intervalo de tiempo determinado es igual al producto de la correspondiente probabilidad de la tasa de reacción y el tiempo transcurrido. Así, por ejemplo, *cada* molécula de mRNA tiene probabilidad $d\delta t$ de degradarse en el intervalo de tiempo $[t, t+\delta t]$.

Por otra parte, la probabilidad de que por ejemplo la reacción de degradación tenga lugar en un intervalo de tiempo δt es proporcional al número de copias de mRNA en t, la tasa de degradación d y el tiempo transcurrido δt. El producto de los dos primeros términos se denominada *propensión* de la reacción (Higham, 2008).

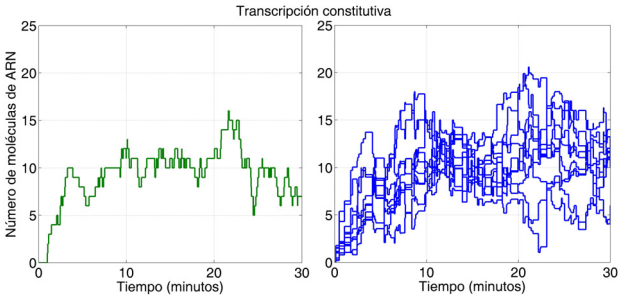

Figura 2: Transcripción constitutiva. Simulación estocástica del número de moléculas de mRNA con el algoritmo SSA. A la izquierda se muestra una realización, y 10 a la derecha. Se usaron valores típicos para *E. coli*: $r = 2{,}5\ \mathrm{min}^{-1}$, $d = 0{,}25\ \mathrm{min}^{-1}$.

Con los elementos arriba explicados, la probabilidad de tener n copias de *mRNA* en el instante $t + \delta t$ es:

$$\begin{aligned} p(n, t+\delta t) = {} & p(n-1,t)r\delta t + p(n+1,t)(n+1)d\delta t + \\ & p(n,t)\left[1 - nd\delta t - r\delta t\right] \end{aligned} \tag{14}$$

Reordenando los términos, y tomando el límite según δt tiende a cero, se llega a la CME expresando la evolución temporal de la distribución de probabilidad:

$$\begin{aligned} \frac{\partial p(n,t)}{\partial t} = {} & d\left[p(n+1,t)(n+1) - p(n,t)n\right] + \\ & r\left[p(n-1,t) - p(n,t)\right] \end{aligned} \tag{15}$$

La ecuación (15) es lineal e infinito dimensional. Hay una ODE para cada posible estado del sistema. Es decir, la CME debe resolverse para todos los posibles valores del número de copias de mRNA, por lo que tendremos:

$$\begin{bmatrix} \frac{\partial p(0,t)}{\partial t} \\ \frac{\partial p(1,t)}{\partial t} \\ \frac{\partial p(2,t)}{\partial t} \\ \vdots \end{bmatrix} = \begin{bmatrix} -r & d & 0 & 0 & 0 & \cdots \\ r & -(r+d) & 2d & 0 & 0 & \cdots \\ 0 & r & -(r+2d) & 3d & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} p(0,t) \\ p(1,t) \\ p(2,t) \\ \vdots \end{bmatrix} \tag{16}$$

4.1. *Simulación estocástica mediante el algoritmo SSA*

Resolver la ecuación (16) es un reto computacional. El algoritmo de Gillespie (o SSA por *Stochastic Simulation Algorithm*) es un algoritmo de simulación de eventos discretos que proporciona soluciones particulares del proceso estocástico que estadísticamente están en correspondencia exacta con los resultados producidos por la CME (Gillespie, 2007).

Las figuras 2 y 3 muestran los resultados obtenidos para las simulaciones estocásticas usando el algoritmo SSA. En la figura 3 se observa que tras el transitorio, la media y la varianza tomadas sobre un gran número de realizaciones del proceso estocástico son iguales. Esto es característico de un proceso de Poisson.

[2] En lo que sigue se abreviará usando *distribuciones de probabilidad* en lugar de *distribuciones de densidad de probabilidad*.

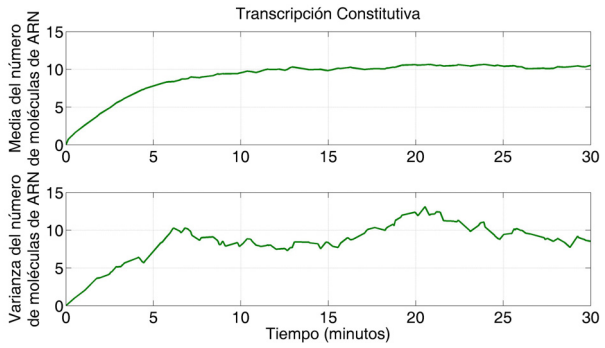

Figura 3: Transcripción constitutiva. Simulación estocástica del número de moléculas de mRNA con el algoritmo SSA. Evolución temporal de la media y la varianza sobre 100 realizaciones del proceso estocástico.

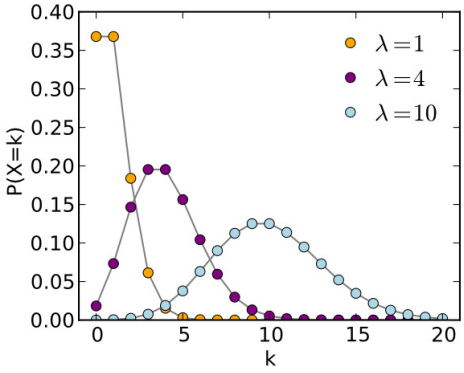

Figura 4: Distribución de Poisson. La distribución de Poisson expresa la probabilidad de que un evento ocurra un número dado de veces en un intervalo fijo de tiempo si estos sucesos ocurren con una ratio media conocida e independientemente del tiempo desde el último suceso.

4.2. *Relación entre el modelo estocástico y el determinista.*

Es interesante comparar el modelo determinista (11) con la evolución del número medio de moléculas proporcionado por el modelo estocástico (15).

Denotando el número medio de moléculas de mRNA como $\langle \text{mRNA} \rangle(t) = \sum_n np(n,t)$, tenemos:

$$\frac{d\langle \text{mRNA} \rangle}{dt} = \sum_n n \frac{\partial p(n,t)}{\partial t} = r - d\langle \text{mRNA} \rangle \qquad (17)$$

Por otro lado, la concentración media $x_{\text{mRNA}} = \frac{\langle \text{mRNA} \rangle}{V_{\text{cell}}}$, donde V_{cell} es el volumen celular. Por tanto, la dinámica determinista de la concentración media coincide con la media de la dinámica estocástica. Es más, se ve que las tasas de reacción probabilísticas y las deterministas coinciden[3], i.e. $r = k_{me}$ y $d = d_m$.

5. Modelo estocástico simplificado: la Ecuación Química de Langevin

La solución analítica de la CME no es posible en el caso general. Aunque existen implementaciones computacionales eficientes, como el algoritmo SSA de Gillespie o las aproximaciones presentadas en (Munsky and Khammash, 2008), las simulaciones estocásticas todavía tienen un alto coste computacional. Por otra parte, la información proporcionada por las aproximaciones analíticas se pierde en gran medida en las aproximaciones basadas en simulaciones. Para superar este contratiempo, se han propuesto varias aproximaciones en tiempo continuo. La Ecuación Química de Langevin (CLE) aproxima la CME por un sistema de ecuaciones diferenciales estocásticas (SDE) de orden igual al número de especies – c.f. con orden igual al posible número de moléculas de todas las especies en la CME (Higham, 2008).

Consideremos de nuevo el modelo de transcripción constitutiva (12), repetido abajo para comodidad del lector.

$$\text{DNA} \xrightarrow{r} \text{DNA} + \text{mRNA}$$

$$\text{mRNA} \xrightarrow{d} \emptyset$$

Asumamos un intervalo de tiempo δt tal que se espera que cada reacción ocurra varias veces en el intérvalo $[t, t+\delta t[$. Bajo estas condiciones, el número de veces que cada reacción se produce será igual a la probabilidad $p(n,t)$ de tener n copias de mRNA en el instante t multiplicada por la tasa probabilística de la reacción y por el intervalo de tiempo δt considerado. Es decir, la frecuencia de reacción es igual a la propensión de la reacción por el intervalo temporal considerado. Esa frecuencia de reacción sigue una distribución de Poisson, con media igual igual a la probabilidad de que la reacción tenga lugar durante δt. La distribución de Poisson dependerá del producto de la propensión por δt, que es usado para parametrizar la distribución. La figura 4 muestra la distribución de Poisson para distintos valores de este parámetro (λ en la figura) [4]. Cuando el valor de λ se incrementa, la distribución de Poisson se puede aproximar por una distribución normal con media λ y varianza λ.

La aproximación de Langevin asume que el parámetro λ para cada reacción, es decir el producto de su propensión por δt, es lo suficientemente grande como para que la función de distribución de probabilidad que da el número de veces que una reacción tiene lugar en el intervalo de tiempo δt, se pueda aproximar por una distribución normal. La distribución de Poisson es discreta – sólo existe para números discretos de realizaciones de cada reacción – mientras que la distribución normal es continua. Esta aproximación implica que el número de veces $T_{ji}(\delta t)$ que tendrá lugar una reacción j afectando a la especie i durante un intervalo de tiempo de duración δt es:

$$T_{ji}(\delta t) = \lambda_j + \sqrt{\lambda_j}\mathcal{N}_j(0,1) \qquad (18)$$

donde $\mathcal{N}_j(0,1)$ es una distribución normal con media cero y varianza unidad[5], y $\mathcal{N}_k(0,1)$ y $\mathcal{N}_l(0,1)$ son independientes pa-

[3] Esto sólo es cierto para reacciones de primer orden, y aproximado para reacciones bimoleculares.

[4] El índice k en el eje horizontal representa el numero de ocurrencias del evento. El parámetro λ es igual al producto de la propensión por δt. Para baja probabilidad de ocurrencia (λ bajo), lo más probable que sólo tenga lugar una ocurrencia del evento en el periodo δt. Para alta probabilidad de ocurrencia (λ alto), lo más probable es que durante el periodo δt, el número de ocurrencias del evento sea igual a la probabilidad media.

[5] De manera que $\sqrt{\lambda_j}\mathcal{N}_j(0,1)$ es una distribución con media cero y varianza λ_j.

ra $k \neq l$, es decir, un ruido blanco gaussiano. Por lo tanto, el número de copias de la especie i de interés en el instante $t + \delta t$ se puede aproximar como la variable aleatoria continua

$$n_i(t + \delta t) = n_i(t) + \sum_j \lambda_j + \sum_j \sqrt{\lambda_j}\mathcal{N}_j(0, 1) \qquad (19)$$

Apliquemos estas ideas al número de moléculas de mRNA en el modelo de transcripción constitutiva (13). En este caso tendremos:

$$n_{\text{mRNA}}(t + \delta t) = n_{\text{mRNA}}(t) + (r - dn_{\text{mRNA}}(t))\delta t + \sqrt{r}\mathcal{N}_1(0, 1)\sqrt{\delta t} - \sqrt{dn_{\text{mRNA}}(t)}\mathcal{N}_2(0, 1)\sqrt{\delta t} \qquad (20)$$

La ecuación (20) corresponde a la discretización Euler-Maruyana de la ecuación diferencial estocástica:

$$\frac{dn_{\text{mRNA}}}{dt} = r - dn_{\text{mRNA}} + \sqrt{r}dW_1 - \sqrt{dn_{\text{mRNA}}}dW_2 \qquad (21)$$

en la que dW_1 y dW_2 son movimientos Brownianos escalares independientes (Higham, 2008). La ecuación (21) se denomina la Ecuación Química de Langevin (CLE), o simplemente la ecuación de Langevin. Los dos primeros términos en el lado derecho de la ecuación corresponden a la cinética determinista. Conjuntamente forman el término macroscópico de deriva. Los términos afectados por el ruido forman el llamado término de difusión. Es interesante notar que el término de deriva determinista crece como la raíz cuadrada del tamaño del sistema. Por lo tanto, el peso relativo del término estocástico con respecto al determinista se escala como la raíz cuadrada inversa del tamaño del sistema. Esto es, según el número de moléculas de las especies se incrementa, la solución de (21) *aproximará* la del modelo determinista en el sentido de que las fluctuaciones alrededor de la solución determinista tendrán un tamaño relativo menor.

6. Tratamiento del ruido en un contexto determinista: la Aproximación Lineal del Ruido

Manejar ecuaciones diferenciales estocásticas es más sencillo que hacerlo con la Ecuación Química Maestra, pero menos que hacerlo con ecuaciones diferenciales ordinarias (ODEs). La Aproximación Lineal del Ruido (LNA) intenta tratar el ruido en un contexto determinista. Básicamente, la LNA ve la trayectoria de la respuesta temporal estocástica de un sistema, en cada instante de tiempo, como la superposición de una respuesta determinista y un término aditivo de ruido, denominado de *fluctuación*. La figura 5 muestra una representación gráfica de la idea usando datos de la simulación estocástica mediante el algoritmo SSA de la sección 4.1. Es muy importante notar que el término de fluctuación no se añade al modelo determinista como una entrada aditiva de ruido, sino que se añade sobre la respuesta temporal del modelo determinista.

Como resultado final de la aplicación del método LNA, se obtiene, por un lado, el modelo determinista que proporciona la trayectoria media del sistema (la concentración media de mRNA en el ejemplo que nos ocupa) y, por otro, una SDE lineal

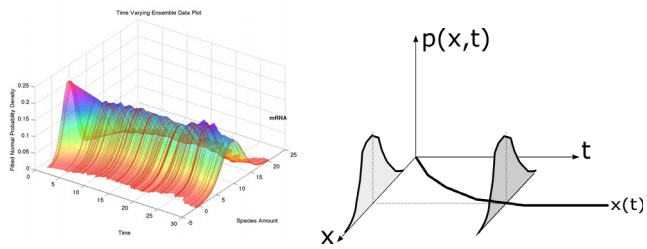

Figura 5: Transcripción constitutiva. **Izq:** Representación 3D de la media y varianza del numero de moléculas de mRNA sobre 100 realizaciones del proceso estocástico obtenidas mediante el algoritmo SSA. **Der.:** Representación de la aproximación LNA.

que proporciona la dinámica de las fluctuaciones. Es importante destacar que la dinámica de la matriz de varianzas-covarianzas del término de fluctuaciones puede obtenerse fácilmente, resultando en una ecuación de tipo Lyapunov. En cada instante de tiempo, la respuesta del sistema se aproxima como la superposición de la respuesta determinista más un término de fluctuación con una matriz de varianzas-covarianzas asociada.

Esta sección sigue los argumentos de (Wallace et al., 2012) para derivar la Aproximación Lineal del Ruido a partir de la discretización de Euler-Maruyana de la CLE obtenida en la sección 5. Así, la CLE (20) correspondiente al número de moléculas de una especie, puede ser escrita de forma general como:

$$n_{\text{mRNA}}(t + \delta t) = n_{\text{mRNA}}(t) + \sum_{m=1}^{M} \nu_m a_m(n_{\text{mRNA}})\delta t + \sum_{m=1}^{M} \nu_m \sqrt{a_m(n_{\text{mRNA}})}\mathcal{N}_m(0, 1)\sqrt{\delta t} \qquad (22)$$

donde:

- M es el número de reacciones. Dos en el ejemplo de transcripción; transcripción efectiva, y degradación de mRNA, que indexaremos como $m = 1, 2$ respectivamente.
- ν_m es el índice estequiométrico de la reacción m. Por ejemplo, la reacción de degradación disminuirá el número de moléculas de mRNA en una. Por tanto, $\nu_2 = -1$. Para el caso de la transcripción efectiva, $\nu_1 = +1$.
- Los términos $a_m(n_{\text{mRNA}}(t))$, con $m = 1 \ldots M$, son las propensiones de las reacciones. Para la transcripción efectiva, asumiendo un número de copias del gen $c_n = 1$, tendremos $a_1(n_{\text{mRNA}}) = rc_n = r$. Para la degradación de mRNA, $a_2(n_{\text{mRNA}}) = dn_{\text{mRNA}}$

Denotemos la concentración de una especie para una realización del proceso estocástico como $x = n/V$, donde n es el número de moléculas, y V el volumen. En lo que sigue estaremos interesados en expresiones usando concentraciones, si bien el desarrollo podría hacerse con número de moléculas. Para ello, expresaremos las propensiones como:

$$\begin{aligned} a_m(n) &= a_m n \quad (\text{e.g. } a_d(n_{\text{mRNA}}) = dn_{\text{mRNA}}) \\ \bar{a}_m(x) &= \bar{a}_m x \quad (\text{e.g. } \bar{a}_d(x_{\text{mRNA}}) = k_d x_{\text{mRNA}}) \end{aligned} \qquad (23)$$

Apoyándonos, tal como vimos en la sección 4, en el hecho de que para reacciones de primer orden las tasas de reacción deterministas y estocásticas coinciden[6] (e.g. $d = k_d$), tendremos que $\bar{a}_m = a_m$. A partir de esta igualdad, obtenemos la relación:

$$a_m(n) = V\bar{a}_m(x) \tag{24}$$

Ahora podemos reescribir la CLE (22) en función de concentraciones:

$$x_{\mathrm{mRNA}}(t+\delta t) = x_{\mathrm{mRNA}}(t) + \sum_{m=1}^{M} \nu_m \bar{a}_m(x_{\mathrm{mRNA}})\delta t + \frac{1}{\sqrt{V}} \sum_{m=1}^{M} \nu_m \sqrt{\bar{a}_m(x_{\mathrm{mRNA}})}\mathcal{N}_m(0,1)\sqrt{\delta t} \tag{25}$$

Obsérvese que conforme aumenta el volumen V, el término estocástico tiende a cero. Sin embargo, que $V \to \infty$ no implica que la concentración tiende a cero. Conforme consideramos más volumen (*tamaño del sistema* en el argot), también consideramos más moléculas, de forma que la concentración se mantiene *casi* constante. Hay que recordar que siempre se consideran intervalos δt muy pequeños, de manera que las variaciones de las variables son muy pequeñas.

Recordemos también que para derivar la CLE en la sección 5 asumimos que cada reacción tiene lugar un número suficiente de veces en el intervalo temporal δt, de manera que el número de veces que tiene lugar la reacción en el intervalo δt puede ser aproximado mediante una distribución normal.

Ahora podemos derivar la LNA como una aproximación a la CLE (25). El punto clave es observar que la CLE difiere de la ecuación determinista en un término proporcional a $1/\sqrt{V}$. Por tanto, buscamos una aproximación de (25) como:

$$x_{\mathrm{mRNA}}(t) = \hat{x}_{\mathrm{mRNA}}(t) + \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t) \tag{26}$$

donde $\xi_{\mathrm{mRNA}}(t)$ es el término de fluctuación, y $\hat{x}_{\mathrm{mRNA}}(t)$ es la solución de la ecuación determinista:

$$\hat{x}_{\mathrm{mRNA}}(t+\delta t) = \hat{x}_{\mathrm{mRNA}}(t) + \sum_{m=1}^{M} \nu_m \bar{a}_m(\hat{x}_{\mathrm{mRNA}})\delta t \tag{27}$$

Obsérvese que (27) es simplemente la discretización de Euler de la ODE correspondiente.

La substitución de (26), y (27) en (25) proporciona:

$$\sum_{m=1}^{M} \nu_m \bar{a}_m(\hat{x}_{\mathrm{mRNA}})\delta t + \frac{1}{\sqrt{V}}\left[\xi_{\mathrm{mRNA}}(t+\delta t) - \xi_{\mathrm{mRNA}}(t)\right] = \sum_{m=1}^{M} \nu_m \bar{a}_m\left(\hat{x}_{\mathrm{mRNA}}(t) + \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t)\right)\delta t + \frac{1}{\sqrt{V}} \sum_{m=1}^{M} \nu_m \sqrt{\bar{a}_m\left(\hat{x}_{\mathrm{mRNA}}(t) + \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t)\right)}\mathcal{N}_m(0,1)\sqrt{\delta t} \tag{28}$$

[6] Esta coincidencia es tanto mayor para reacciones de orden superior cuanto mayor número de moléculas aumenta.

Si estamos próximos a la región en la que el modelo determinista es válido, entonces $\hat{x}_{\mathrm{mRNA}}(t) \approx x_{\mathrm{mRNA}}(t)$. En otras palabras $\frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t)$ es suficientemente pequeño como para aproximar por la serie Taylor truncada:

$$\bar{a}_m\left(\hat{x}_{\mathrm{mRNA}}(t) + \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t)\right) = \bar{a}_m\left(\hat{x}_{\mathrm{mRNA}}(t)\right) + \left.\frac{\partial \bar{a}_m(x)}{\partial x}\right|_{x=\hat{x}_{\mathrm{mRNA}}(t)} \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t) \tag{29}$$

Nota: En (29) estamos haciendo uso implícito del hecho que tenemos un sistema escalar. En el caso general de un sistema con N especies y M reacciones, el Jacobiano será una matriz. Del mismo modo, el término de perturbación ξ será un vector.

La substitución de (29) en (28) proporciona, despreciando el término en $V^{-\frac{1}{2}}$ que aparece dentro de la raíz cuadrada:

$$\xi_{\mathrm{mRNA}}(t+\delta t) = \xi_{\mathrm{mRNA}}(t) + \sum_{m=1}^{M} \nu_m \left.\frac{\partial \bar{a}_m(x)}{\partial x}\right|_{x=\hat{x}_{\mathrm{mRNA}}(t)} \xi_{\mathrm{mRNA}}(t)\delta t + \sum_{m=1}^{M} \nu_m \sqrt{\bar{a}_m\left(\hat{x}_{\mathrm{mRNA}}(t)\right)}\mathcal{N}_m(0,1)\sqrt{\delta t} \tag{30}$$

La ecuación (30) proporciona la dinámica de las fluctuaciones alrededor de la solución determinista. Se trata de una SDE, pero de una SDE lineal variante en el tiempo, de la forma $d\xi = A(t)\xi + B(t)d\omega$. Esto es importante, puesto que se trata de una ecuación que podemos resolver analíticamente.

En el caso general, tenemos un sistema con N especies químicas y M reacciones. Llamemos $\bar{a}(\hat{x}) \in \mathbb{R}^M$ al vector de propensiones de reacción, $S = [\nu_{ij}] \in \mathbb{R}^{N\times M}$ a la matriz estoquiométrica, $\hat{x} \in \mathbb{R}^N$ a las concentraciones de las especies dadas por el modelo determinista, $\xi(t) \in \mathbb{R}^N$ al vector de fluctuaciones, $J = \left.\frac{\partial \bar{a}(x)}{\partial x}\right|_{x=\hat{x}} \in \mathbb{R}^{M\times N}$ es la matriz Jacobiana del vector de propensiones, y $\mathrm{diag}\left(\sqrt{\bar{a}}\right)$ la matriz diagonal conteniendo la raíz cuadrada de las propensiones $\bar{a}(\hat{x})$ en la diagonal y ceros en el resto de posiciones. Entonces, el modelo determinista en el caso general se puede expresar como:

$$\dot{\hat{x}}(t) = S\,\bar{a}(\hat{x}) \tag{31}$$

y la dinámica discretizada de la LNA para el término de fluctuación como:

$$\xi(t+\delta t) = \xi(t) + S\,J\xi(t)\delta t + S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}(0,1)\sqrt{\delta t} \tag{32}$$

donde $\mathcal{N}(0,1) \in \mathbb{R}^M$ es un vector con M ruidos gaussianos independientes de media cero y varianza unidad. Se recuerda también que en la LNA la concentración de las especies $x(t)$ se aproxima como:

$$x(t) = \hat{x}(t) + \frac{1}{\sqrt{V}}\xi(t) \tag{33}$$

donde V es el volumen.

La ecuación (32) puede resolverse numéricamente. En la práctica, sin embargo, estamos más interesados en las características estadísticas del término de fluctuación $\xi(t)$. Afortunadamente, la dinámica de la matriz de varianzas-covarianzas

de $\xi(t)$ puede obtenerse fácilmente, tal como se muestra en el Apéndice˜A. Denotando por $C(t) \in \mathbb{R}^{N\times N}$ la citada matriz de varianzas-covarianzas del término de fluctuación, la dinámica de $C(t)$ responde a la ecuación de tipo Lyapunov:

$$\dot{C} = CJ^T S^T + SJC + S\,\mathrm{diag}(\bar{a}(\hat{x}))\,S^T \tag{34}$$

Obsérvese que esta ecuación está acoplada a la dinámica determinista (31) ya que el valor de las propensiones $\bar{a}$, y su Jacobiano J serán, en el caso general, una función del estado del sistema $\hat{x}$.

6.1. Simulación de la LNA

A continuación, aplicamos estos resultados a nuestro ejemplo de transcripción. El modelo determinista, usando la notación anterior, se puede expresar como:

$$\dot{\hat{x}}_{\mathrm{mRNA}} = S\,\bar{a}(\hat{x}) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} k_r c_n \\ k_d \hat{x}_{\mathrm{mRNA}} \end{bmatrix} \tag{35}$$

El Jacobiano del vector de propensiones es:

$$J = \left.\frac{\partial \bar{a}(x)}{\partial x}\right|_{x=\hat{x}} = \begin{bmatrix} 0 \\ k_d \end{bmatrix} \tag{36}$$

A partir de ellos se puede construir el sistema extendido formado por la dinámica determinista de la concentración media de mRNA, y la de la matriz de varianzas-covarianzas de la fluctuación. En nuestro ejemplo escalar se tendrá:

$$\begin{aligned} \dot{\hat{x}}_{\mathrm{mRNA}} &= k_r c_n - k_d \hat{x}_{\mathrm{mRNA}} \\ \dot{C} &= -2k_d C + (k_r c_n + k_d \hat{x}_{\mathrm{mRNA}}) \end{aligned} \tag{37}$$

Una aproximación útil se obtiene usando el valor de régimen permanente de la concentración de mRNA, $\hat{x}_{\mathrm{mRNA}ss} = k_r c_n / k_d$, en la ecuación de la dinámica de la varianza:

$$\begin{aligned} \dot{\hat{x}}_{\mathrm{mRNA}} &= k_r c_n - k_d \hat{x}_{\mathrm{mRNA}} \\ \dot{C} &= -2k_d C + 2k_r c_n \end{aligned} \tag{38}$$

En (38) la dinámica de C está desacoplada de la de $\hat{x}_{\mathrm{mRNA}}$. Por tanto, la varianza puede obtenerse independientemente, lo cual es de utilidad en el análisis de circuitos genéticos complejos. La figura 6 muestra los resultados obtenidos en ambos casos en unidades de concentracion (eje izquierdo).

Podemos expresar los resultados de la LNA en función de número de moléculas. Como puede verse en los ejes de la derecha de la figura 6, se obtiene un valor de régimen permanente de $\hat{n}_{\mathrm{mRNA}} = 10$, y el mismo valor de varianza, como corresponde a un proceso de Poisson. Se trata de los mismos valores que se obtuvieron mediante la CME y la CLE.

Hay que recordar que la concentración *real* x_{mRNA} se obtendría como

$$x_{\mathrm{mRNA}}(t) = \hat{x}_{\mathrm{mRNA}}(t) + \frac{1}{\sqrt{V}}\xi_{\mathrm{mRNA}}(t) \tag{39}$$

donde $\xi_{\mathrm{mRNA}}(t)$ es el término de fluctuación con varianza $C(t)$.

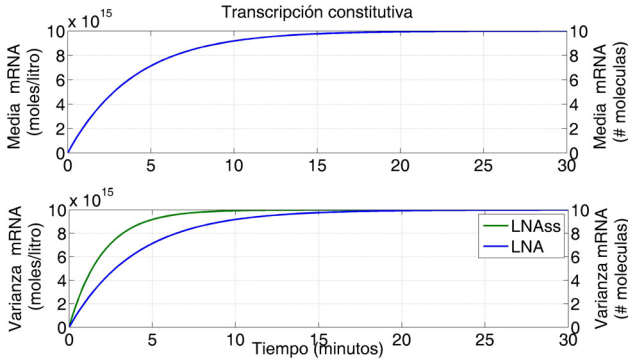

Figura 6: **Arr.:** En el eje de la izquierda, la concentración media de mRNA $\hat{x}_{\mathrm{mRNA}}$. Unidades en moléculas/litro. En el eje de la derecha, el número medio de moléculas de mRNA $\hat{n}_{\mathrm{mRNA}}$. **Bajo:** Comparación entre la varianza de $\xi(t)$ obtenida usando (37), y (38). Eje izq: en concentraciones. Eje dcha: Unidades en número de moléculas al cuadrado por célula.

6.2. Magnitud del estado.

Es interesante notar la relación entre la ecuación (32) y los sistemas estocásticos típicamente encontrados en la literatura de control automático. Estos suelen responder a una SDE lineal, en la cual el término de ruido no depende del estado (Glad and Ljung, 2000; Aström, 2006), como:

$$x(t+\delta t) = x(t) + Ax(t)\delta t + S\Lambda^{\frac{1}{2}}\mathcal{N}(0,1)\sqrt{\delta t} \tag{40}$$

donde $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{N\times N}$, $S \in \mathbb{R}^{N\times N}$, Λ es una matriz diagonal con componentes $\Lambda_{ii} = \lambda_i\, i = 1 \ldots N$, y $\mathcal{N}(0,1) \in \mathbb{R}^N$ es un vector con N señales de ruido gaussiano independientes, de media cero y varianza unidad[7].

La dinámica de la matriz de varianzas-covarianzas del vector de estado x sigue una dinámica tipo Lyapunov completamente análoga a la de 34.

Denotemos por $\mu \triangleq \langle x\rangle(t)$ la media sobre un conjunto de (infinitas) realizaciones del proceso estocástico (40), $\Pi(t) \triangleq \langle xx^T\rangle(t)$, y $\Xi(t) \triangleq \Pi - \mu\mu^T$ su matriz de varianzas-covarianzas. Nótese que hablamos de media obtenida en cada t sobre muchas realizaciones del proceso, no sobre el tiempo para una realización. Calculando la media en (40) tendremos:

$$\langle x\rangle(t+\delta t) = \langle x\rangle(t) + A\langle x\rangle(t)\delta t \tag{41}$$

donde el último término se anula porque el vector de ruidos gausianos $\mathcal{N}(0,1)$ tiene media cero para todos sus componentes. Por tanto, conforme δt tiende a cero, obtenemos el modelo determinista medio, definido por la ODE:

$$\dot{\mu}(t) = A\mu(t) \tag{42}$$

Ahora podemos obtener la dinámica de la matriz $\Pi(t) =< xx^T >$

[7] De modo que $\sqrt{\lambda_i}\mathcal{N}_i(0,1)$ es un ruido gaussiano de media cero, y varianza λ_i.

(t) siguiendo pasos análogos a los del Apéndice˜A. Así

$$\begin{aligned} xx^T(t+\delta t) =& xx^T(t) + \left(xx^T A^T + Axx^T\right)\delta t + Axx^T A^T(\delta t)^2 \\ &+ \left(x\mathcal{N}^T\Lambda^{\frac{1}{2}}S^T + S\Lambda^{\frac{1}{2}}\mathcal{N}x^T\right)\sqrt{\delta t} \\ &+ \left(Ax\mathcal{N}^T\Lambda^{\frac{1}{2}}S^T + S\Lambda^{\frac{1}{2}}\mathcal{N}x^T A^T\right)(\delta t)^{\frac{3}{2}} \\ &+ S\Lambda^{\frac{1}{2}}\mathcal{N}\mathcal{N}^T\Lambda^{\frac{1}{2}}S^T\delta t \end{aligned} \tag{43}$$

Puesto que δt es muy pequeño, podemos despreciar los términos en $(\delta t)^{\frac{3}{2}}$ y $(\delta t)^2$. Por otro lado, haremos uso del hecho que $\langle\mathcal{N}\mathcal{N}^T\rangle = I_{N\times N}$ ya que $\mathcal{N}$ son señales de ruido independientes de varianza unidad, y $\langle\mathcal{N}x^T\rangle = \langle x\mathcal{N}^T\rangle = 0_{N\times N}$ for $\mathcal{N}$ y x son independientes. Tomando medias sobre realizaciones en(43), obtenemos:

$$\Pi(t+\delta t) = \Pi(t) + \left(\Pi A^T + A\Pi\right)\delta t + S\Lambda S^T\delta t \tag{44}$$

Finalmente, tomando límite cuando δt tiende a cero, obtenemos la ODE correspondiente a la dinámica de la *magnitud del estado* Π:

$$\dot{\Pi} = \Pi A^T + A\Pi + S\Lambda S^T \tag{45}$$

La dinámica de la matriz de varianzas-covarianzas $\Xi(t) \triangleq \Pi - \mu\mu^T$, se obtiene a partir de (42) y (45):

$$\begin{aligned} \dot{\Xi} &= \left(\Xi + \mu\mu^T\right)A^T + A\left(\Xi + \mu\mu^T\right) + S\Lambda S^T - A\mu\mu^T - \mu\mu^T A^T \\ &= \Xi A^T + A\Xi + S\Lambda S^T \end{aligned} \tag{46}$$

Como puede verse, se trata de un caso particular del tratado en la sección 6.

7. Conclusión

En este tutorial hemos visto algunas de las metodologías utilizadas habitualmente en el modelado de circuitos biológicos característicos de la biología molecular celular. El ejemplo conductor utilizado, la transcripción genética no regulada, es lo suficientemente simple como para poder ver la aplicación de estas metodologías con un cierto detalle. Obviamente, se paga el precio de perder la visión de algunas de las dificultades que surgen en el modelado de circuitos genéticos más complejos. La Biología de Sistemas, y la Biología Sintética, utilizan una jerarquía de circuitos genéticos de complejidad creciente. En la actualidad existen grandes retos en el modelado y simulación de circuitos biológicos.

Todavía hacen falta técnicas para la construcción sistemática de modelos de gran escala –en el límite, escala genómica (Chakrabarti et al., 2013)– y con fuerte presencia de ruido e incertidumbre. Las practicas actuales de construcción de modelos en biología aun ignoran en gran medida pasos cruciales como el análisis de identifiabilidad, el diseño optimo de experimentos, y la cuantificación de la incertidumbre, aspecto este último de vital importancia en aplicaciones biológicas (Villaverde and Banga, 2014; Kiparissides et al., 2011).

La modularidad, esto es el uso de módulos caracterizados y su interconexión, es una de las características fundamentales de los sistemas diseñados artificialmente. La caracterización aislada de módulos biológicos no es sencilla, ya que su comportamiento depende del contexto en el que se integran. En efecto, entre los módulos hay intercambio de materia, dando lugar a efectos de carga. En este sentido, la aproximación al diseño modular es menos parecida al típico diagrama de bloques independientes de la automática, y más parecida al diseño en ingeniería eléctrica y civil, con el uso de optimización y caracterización de cargas e interacciones entre módulos (Church et al., 2014). Recientemente han aparecido propuestas, basadas en el uso del método de perturbaciones singulares, para la caracterización de la carga inducida por la conexión entre módulos biológicos, denominada *retroactividad* en el campo (Jayanthi et al., 2013; Del Vecchio, 2013).

La simulación de circuitos biológicos complejos mediante los algoritmos vistos en el tutorial tiene un alto coste computacional. Este se incrementa cuando se consideran poblaciones celulares en las que los individuos intercambian información vía comunicación celular. En el tutorial hemos visto que las propensiones responden a reacciones primitivas (i.e. reacciones de enlace, degradación, etc.). Recientemente se ha demostrado que pueden derivarse modelos tipo CLE y LNA derivados directamente a partir de la reducción de modelos deterministas complejos. En estos casos las propensiones toman la forma de funciones complejas de tipo polinomial racional (Boada et al., 2015). No obstante, la obtención de estos modelos sigue adoleciendo de falta de metodologías sistemáticas. Por último, además de estas aproximaciones, hacen falta algoritmos de simulación tipo SSA más eficientes.

En resumen, el campo de la biología, en sus facetas de Biología de Sistemas y Biología Sintética, ofrece muchos retos en los que la ingeniería de sistemas y automática pueden aportar soluciones. Este tutorial se ha centrado en aspectos de modelado de circuitos a nivel celular. La caracterización robusta de los mismos, teniendo la incertidumbre y estocasticidad inherentes, así como el diseño y de dispositivos biológicos y sus mecanismos de regulación son áreas que ofrecen un gran potencial.

English Summary

Modelling biochemical systems: from Mass Action Kinetics to Linear Noise Approximation.

Abstract

In the last decade we have witnessed a growing application of engineering techniques to biology. Areas such as Systems Biology or, more recently, Synthetic Biology, get more and more attention from the engineers. Specifically, modeling in these fields makes possible the generation of new experimentally verifiable hypothesis, and new ways of biological intervention, as well as more or less mechanistic explanations of experimental results. A model-based approximation requires the consideration of the biochemical reactions dynamics and their regulation. The first part of this tutorial describes the deterministic modeling and model reduction techniques, as applied to the class of biochemical reactions specific to molecular cell biology.

Noise plays a crucial role in the biological circuitry dynamics. In the field of automatic control there is a long tradition of modeling using linear stochastic differential equations, under the simplifying assumption that noise has a magnitude independent of the state. This assumption is not valid in biological circuits. The second part of the tutorial describes the most widely used methods for stochastic modeling in molecular cell biology, paying special attention to the so-called linear noise approximation.

Keywords:

Stochastic systems, Differential equations, Modeling of continuous systems, Model reduction, Simulation, Noise, Biological and biotechnological systems and bioprocesses.

Agradecimientos

Este trabajo ha sido realizado parcialmente gracias al apoyo de los proyectos FEDER-CICYT DPI2011-28112-C04-01, y DPI2014-55276-C5-1. Yadira Boada agradece la beca FPI/2013-3242 de la Universitat Politècnica de València.

Referencias

Alberts, B., Bray, D., Hopkin, K., Johnson, A. D., Johnson, A., Roberts, K., Lewis, J., Raff, M., Walter, P., 2009. Essential Cell Biology, 3rd Edition. Garland Science.

Arpino, J. A. J., Hancock, E. J., Anderson, J., Barahona, M., Stan, G.-B. V. B., Papachristodoulou, A., Polizzi, K., 7 2013. Tuning the dials of synthetic biology. Microbiology 159 (Pt 7), 1236–53.

Aström, K. J., 2006. Introduction to Stochastic Control Theory. Dover.

Blanchini, F., Franco, E., 2011. Structurally robust biological networks. BMC Systems Biology 5 (1), 74.

Boada, Y., Vignoni, A., Navarro, J. L., Picó, J., 2015. Improvement of a cle stochastic simulation of gene synthetic network with quorum sensing and feedback in a cell population. In: Proceedings ECC 15.

Chakrabarti, A., Miskovic, L., Soh, K., Hatzimanikatis, V., 2013. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotech. J. 8 (9), 1043.

Chellaboina, V., Bhat, S., Haddad, W., Bernstein, D., 8 2009. Modeling and analysis of mass-action kinetics. IEEE Control Systems Magazine 29 (4), 60–78.

Chen, W. W., Niepel, M., Sorger, P. K., 2010. Classic and contemporary aproaches to modeling biochemical reactions. Genes & development 24, 1861–1875.

Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. A., Weiss, R., 4 2014. Realizing the potential of synthetic biology. Nat Rev Mol Cell Biol 15 (4), 289–94.

De Lorenzo, V., 8 2014. Biología sintética: la ingeniería al asalto de la complejidad biológica. Arbor 190 (768), a149.

Del Vecchio, D., 2013. A control theoretic framework for modular analysis and design of biomolecular networks. Annual Reviews in Control 37 (2), 333–345.

ERASynBio, 2014. Next steps for european synthetic biology: a strategic vision. Tech. rep., ERASynBio.
URL: `https://www.erasynbio.eu`

Gillespie, D. T., 2007. Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58, 35–55.

Glad, T., Ljung, L., 2000. Control Theory. Multivariable and Nonlinear Methods. Taylor & Francis.

Higham, D. J., 1 2008. Modeling and simulating chemical reactions. SIAM Review 50 (2), 347–368.

Hinch, E. J., 1991. Perturbation Methods. Cambridge Texts in Applied Mathematics. Cambridge U. Press.

Jayanthi, S., Nilgiriwala, K. S., Del Vecchio, D., 2013. Retroactivity controls the temporal dynamics of gene transcription. ACS synthetic biology.

Khalil, H. K., 2011. The Control Handbook. CRC Press, Ch. Two Timescale and Averaging Methods.

Kiparissides, A., Koutinas, M., Kontoravdi, C., Mantalaris, A., Pistikopoulos, E. N., 2011. Closing the loop in biological systems modeling: From the in silico to the in vitro. Automatica 47, 1147–1155.

Kokotovic, P., Khalil, H., O'Reilly, J., 1986. Singular perturbation methods in control: analysis and design. Academic Press.

Kwok, R., 2010. Five hard truths for synthetic biology. Nature 463, 288–290.

Llaneras, F., Picó, J., 1 2008. Stoichiometric modelling of cell metabolism. J Biosci Bioeng 105 (1), 1–11.

Mélykúti, B., Hespanha, J. P., Khammash, M., 2014. Equilibrium distributions of simple biochemical reaction systems for time-scale separation in stochastic reaction networks. Journal of The Royal Society Interface 11 (97), 20140054.

Milo, R., Jorgensen, P., Moran, U., Weber, G., Springer, M., 1 2010. Bionumbers–the database of key numbers in molecular and cell biology. Nucleic Acids Res 38 (Database issue), D750–3.

Munsky, B., Khammash, M., 2008. The finite state projection approach for the analysis of stochastic noise in gene networks. Automatic Control, IEEE Transactions on 53 (Special Issue), 201–214.

Picó-Marco, E., 2013. Differential algebra for control systems design. computation of canonical forms. Control Systems Magazine 33 (2), 52–62.

Scott, M., Hwa, T., Ingalls, B., 5 2007. Deterministic characterization of stochastic genetic circuits. Proc Natl Acad Sci U S A 104 (18), 7402–7.

Scott, M., Ingalls, B., Kaern, M., 6 2006. Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. Chaos 16 (2), 026107.

Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M., Wishart, D. S., 1 2004. The cybercell database (ccdb): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of escherichia coli. Nucleic Acids Res 32 (Database issue), D293–5.

Villaverde, A. F., Banga, J. R., 2014. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J. Royal Soc. Interface 11:20130505.

Wallace, E., Gillespie, D., Sanft, K., Petzold, L., 2012. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. IET Systems Biology 6 (4), 102–115.

Zagaris, A., Kaper, H. G., Kaper, T. J., 1 2004. Analysis of the computational singular perturbation reduction method for chemical kinetics. Journal of Nonlinear Science 14 (1), 59–91.

Apéndice A. Solución de la Ecuación Lineal del Ruido

Denotemos por $\langle\xi\rangle(t)$ la media sobre un conjunto (infinito) de soluciones particulares del proceso estocástico (32), y por $C(t) = \langle\xi\xi^T\rangle(t)$ su matriz de varianzas-covarianzas.

Asumamos que en $t = t_0$ empezamos desde una condición inicial precisa $\hat{x}(t_0) = x(t_0)$. Esto es, $\xi(t_0) = 0$, y $\langle\xi\rangle(t_0) = 0$ por que no hay incertidumbre en el estado inicial. Esencialmente, la aleatoriedad surgirá con el tiempo, desde un estado inicial dado cierto, debido a la estocasticidad intrínseca del proceso. Seguidamente veremos que $\langle\xi\rangle(t) = 0$ para todo $t \geq 0$. Esto es, la media en cada instante de tiempo sobre múltiples realizaciones particulares del proceso estocástico, es cero para todos los N componentes del término de fluctuaciones ξ. Para verlo, se parte de la media en la ecuación (32). Entonces:

$$\langle\xi\rangle(t + \delta t) = \langle\xi\rangle(t) + S\,J\langle\xi\rangle(t)\delta t \tag{A.1}$$

donde el último término desaparece porque el término de ruido gaussiano $\mathcal{N}(0, 1)$ tiene media cero para todos sus componentes. Consecuentemente, según δt va a cero, podemos establecer

la ODE:

$$\langle\dot{\xi}\rangle(t) = S\,J\langle\xi\rangle(t) \tag{A.2}$$

Si, como se ha asumido, $\langle\xi\rangle(t_0) = 0$, entonces $\langle\xi\rangle(t) \equiv 0$ es una solución de (A.2). Como prueba alternativa, se puede proceder simplemente por inducción usando la actualización en tiempo discreto (A.1).

Ahora obtendremos la dinámica de la matriz de varianzas-covarianzas $C(t) = \langle\xi\xi^T\rangle(t)$. Con este fin, primero tomaremos la ecuación (32), y multiplicaremos $\xi(t+\delta t)$ por su traspuesta. Seguidamente, tomaremos la media sobre el resultado. Con lo cual:

$$\begin{aligned}\xi\xi^T(t+\delta t) = \xi\xi^T(t)+&\\ &\left(\xi\xi^T J^T S^T + S\,J\xi\xi^T\right)\delta t + S\,J\xi\xi^T J^T S^T(\delta t)^2+\\ &\left(\xi\mathcal{N}^T\mathrm{diag}\left(\sqrt{\bar{a}}\right)S^T + S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}\xi^T\right)\sqrt{\delta t}+\\ &\left(S\,J\xi\mathcal{N}^T\mathrm{diag}\left(\sqrt{\bar{a}}\right)S^T + S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}\xi^T J^T S^T\right)(\delta t)^{\frac{3}{2}}\\ &+ S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}\mathcal{N}^T\mathrm{diag}\left(\sqrt{\bar{a}}\right)S^T\delta t\end{aligned} \tag{A.3}$$

Dado que δt es muy pequeño, podemos despreciar los términos en $(\delta t)^{\frac{3}{2}}$ y $(\delta t)^2$, obteniendo:

$$\begin{aligned}\xi\xi^T(t+\delta t) =&\xi\xi^T(t) + \left(\xi\xi^T J^T S^T + S\,J\xi\xi^T\right)\delta t+\\ &\left(\xi\mathcal{N}^T\mathrm{diag}\left(\sqrt{\bar{a}}\right) + S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}\xi^T\right)\sqrt{\delta t}+\\ &S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathcal{N}\mathcal{N}^T\mathrm{diag}\left(\sqrt{\bar{a}}\right)S^T\delta t\end{aligned} \tag{A.4}$$

Ahora podemos tomar la media sobre las realizaciones particulares en la ecuación (A.4). Podemos tomar la media sobre múltiples realizaciones particulares del experimento, no en el tiempo sobre una única solución particular del mismo. Sólo en el caso de procesos ergódicos (débilmente) estacionarios ambos cálculos darán el mismo resultado. Haremos uso del hecho que $\langle\mathcal{N}\mathcal{N}^T\rangle = I_{N\times N}$ para $\mathcal{N}$ son señales de ruido independientes de varianza uno, y $\langle\mathcal{N}\xi^T\rangle = \langle\xi\mathcal{N}^T\rangle = 0_{N\times N}$ para $\mathcal{N}$ y ξ son independientes. Además se recuerda que se definió $C(t) = \langle\xi\xi^T\rangle(t)$. Entonces:

$$C(t+\delta t) = C(t)+\left(CJ^T S^T + S\,JC\right)\delta t+S\,\mathrm{diag}\left(\sqrt{\bar{a}}\right)\mathrm{diag}\left(\sqrt{\bar{a}}\right)S^T\delta t \tag{A.5}$$

Tomando el límite cuando δt tiende a cero, finalmente obtenemos la ODE que define la dinámica de la matriz de varianzas-covarianzas de las fluctuaciones:

$$\dot{C} = CJ^T S^T + S\,JC + S\,\mathrm{diag}(\bar{a}(\hat{x}))\,S^T \tag{A.6}$$

Nótese que esta ecuación está acoplada con la dinámica determinista (31) por que el valor de las propensidades $\bar{a}$, y su Jacobiano J serán, en el caso general, una función del estado del sistema $\hat{x}$.

Si sólo estamos interesados en la matriz de varianzas-covarianzas en régimen permanente, la ecuación (A.6) puede resolverse en régimen permanente, resultando un sistema de ecuaciones lineales algebraicas.

Una simplificación alternativa viene de considerar la respuesta determinista en régimen permanente cuando se obtienen tanto el Jacobiano de las propensiones J como la matriz diagonal $\mathrm{diag}(\bar{a})$ (Scott et al., 2006, 2007). En tal caso, tenemos

$$\begin{aligned}0 &= S\,\bar{a}(\hat{x}_{ss})\\ J_{ss} &= \left.\frac{\partial\bar{a}(x)}{\partial x}\right|_{x=\hat{x}_{ss}}\\ \dot{C} &= CJ_{ss}^T S^T + S\,J_{ss}C + S\,\mathrm{diag}(\bar{a}(\hat{x}_{ss}))\,S^T\end{aligned} \tag{A.7}$$

en este caso, la dinámica de C puede obtenerse *a priori*, independientemente de la solución de la trayectoria determinista.

En algunas ocasiones puede ser interesante tener la dinámica de la matriz de varianzas-covarianzas cuando la aproximación (33) se expresa en términos del número de moléculas en vez de las concentraciones. Multiplicando ambos lados de la ecuación (33) por el volumen V:

$$n(t) = \hat{n}(t) + \sqrt{V}\xi(t) \tag{A.8}$$

Definamos $\bar{\xi}(t) \triangleq \sqrt{V}\xi(t)$, y $\bar{C}(t) \triangleq \langle\bar{\xi}\bar{\xi}^T\rangle(t)$ la correspondiente matriz de varianzas-covarianzas. Entonces, claramente $\bar{C}(t) = VC(t)$. Tomando la derivada, y usando la ecuación (A.6), y el hecho de que las propensiones en el número de moléculas y concentraciones están relacionadas por $a(\hat{n}) = V\bar{a}(\hat{x})$, obtenemos:

$$\dot{\bar{C}} = \bar{C}J^T S^T + S\,J\bar{C} + S\,\mathrm{diag}(a(\hat{n}))\,S^T \tag{A.9}$$

Por lo tanto, para obtener la matriz de varianzas-covarianzas en términos del número de moléculas sólo hay que resolver la ecuación (A.9), que es idéntica a (A.6) excepto por el uso de propensiones expresadas usando el número de moléculas.

Además, a la vista de la ecuación (33), en ocasiones es útil tener la varianza-covarianza $\tilde{C}(t)$ para el término $\tilde{\xi}(t) \triangleq \frac{1}{\sqrt{V}}\xi(t)$. En este caso, usando $\tilde{C}(t) = C(t)/V$ se obtiene:

$$\dot{\tilde{C}} = \tilde{C}J^T S^T + S\,J\tilde{C} + \frac{1}{V}S\,\mathrm{diag}(\bar{a}(\hat{x}))\,S^T \tag{A.10}$$