Criticalidad, Robustez y Evolucionabilidad en Redes de Regulación Genética

Enrique Balleza Dávila

30 de octubre de 2007

.

¡Qué extranjero!

Yo veo puro cine del séptimo arte.

Mi moto alpina derrapante.

¡Te vas a arruinar tus manos mamacita!

¡Hasta le salio pelo al Kenkre!

Su lechita y a dormir.

No, esa si no la voy a ver Max.

Yo era un chico muy enojado...

Claro que ésta motivado, ¡si vamos a tener gemelitos!

Arqueología urbana: *¡Vamos al chorrito!*

Y el Enrique ahí bien atento escuchando al Kauffman mientras yo me andaba asfixiando.

.

Todo fin puede determinarse por un fin superior con respecto al cual aquel se convierte en medio. De aquí nace una jerarquía de fines y medios que debe poseer un elemento último: el bien.

— **Aristóteles**

Un producto organizado es aquel en el que todo es, recíprocamente, medio y fin.

— **Kant**

Índice general

Resumen

Usando el modelo de Kauffman, mostramos que redes reales de regulación operan alrededor de la fase crítica lo que otorga la clara ventaja evolutiva de adaptación y, simultáneamente, estabilidad. Para mostrar lo anterior, implementamos modelos Booleanos de redes reales y caracterizamos su dinámica global. Realizamos la implementación Booleana usando las redes que se conocen de tres microorganismos (*E. coli*, *B. subtilis* y *S. cerevisiae*) y cientos de microarreglos. A partir de estos datos, calculamos el valor más probable de la probabilidad de expresión genética, p, con inferencia paramétrica Bayesiana. Implementamos el modelo dinámico de las redes, proponiendo conjuntos de funciones aleatorias acordes con el valor de p encontrado. La caracterización dinámica se lleva a cabo al medir la distancia de Hamming entre configuraciones similares cuando éstas evolucionan en la red. Hay tres casos: cuando la dinámica es caótica la distancia entre configuraciones aumenta, cuando es ordenada disminuye. Finalmente, cuando es crítica se conserva. El último caso se encuentra en la transición entre orden y caos. Los resultados indican que las redes reales se encuentran alrededor de la transición. Lo mismo sucede en la red de regulación del desarrollo floral de la planta *A. thaliana* y en la red de regulación que controla la formación de los segmentos polares en *D. melanogaster*, cuyas funciones y topología se conocen completamente. La ubicuidad de la criticalidad en organismos tan dispares hace pensar que esta importante característica dinámica es genérica.

Por otra parte –suponiendo que el modelo de Kauffman captura las propiedades esenciales de la regulación–, mostramos que las redes, bajo la perturbación de duplicación seguida de divergencia genética, son robustas y evolucionables. La perturbación consiste en añadir un nuevo gen a la red. Esto provoca que el paisaje de atractores se transforme. En particular, los atractores se fusionan, dividen, ganan configuraciones, las pierden, etc. pero también pueden quedar idénticos, sucediendo esto con mayor probabilidad en redes ordenadas y críticas. Además de los cambios anteriores, también pueden surgir nuevos atractores o desaparecer completamente. La conservación de los atractores inicialmente codificados por la red se identifica con el concepto de robustez; el surgimiento de nuevos con el de evolucionabilidad. La robustez y evolucionabilidad son máximas en redes críticas. Creemos que estos resultados muestran que la estabilización de configuraciones transientes en nuevos atractores –sin que esto implique la pérdida de los ya codificados– es una propiedad esencial de las redes de regulación.

Abstract

Using the Kauffman model, we show that real gene regulatory networks operate near a critical dynamic phase. Criticality confers to the network the clear evolutionary advantage of adaptability without loosing stability. To characterize the global dynamics of real gene regulatory networks we implement Boolean models. We do this for three microorganisms (*E. coli*, *B. subtilis* and *S. cerevisiae* using their known regulatory networks and hundreds of microarray experiments. With these data and using Bayesian parametric inference we calculate the most probable value of the probability of gene expression p for each organism. We use the calculated p values to generate sets of Boolean functions and implement the dynamical Boolean model of real networks. We characterize network dynamics measuring the evolution of the Hamming distance between similar network configurations when they evolve one time step. Three possibilities arise. When dynamics are chaotic the distance between configurations grow, when dynamics are ordered the distance diminishes. Finally, the distance stays the same when dynamics are critical. This last case lies in the transition between order and chaos. The results indicate that real networks operate around this transition. This same happens in the regulatory network of cell fate determination of flower development of *A. thaliana* and in the regulatory network of genes responsible for the polarity segmentation of *D. melanogaster*. For this networks the Boolean functions and connections are completely known. The ubiquity of criticality in networks of organisms spanning four kingdoms of life suggests that this property may be a generic characteristic of life.

Using numerical simulations, we also show that gene regulatory networks perturbed by gene duplication followed by gene divergence are robust and evolvable. The perturbation modifies the attractor landscape. In particular, attractors coalesce, divide, gain configurations, loose configurations, etc. but also there is a chance that they remain the same. This happens with high probability in ordered and critical networks. There is also a possibility for new attractors to emerge after the perturbation or to disappear completely. The conservation of the attractors initially codified by the unperturbed network is identified with the concept of biological robustness; the emerge of new attractors with evolvability. Robustness and evolvability are maximal for critical networks. These results show that the stabilization of transient configurations into attractors –without loosing attractors previously codified by the network– is an essential property of gene regulatory networks.

Capítulo 1

Introducción

En años recientes la biología celular ha experimentado avances espectaculares en el control que se tiene de la maquinaria de expresión y en el sondeo global de muchas características importantes. En cuanto al control, actualmente es posible crear *switches* genéticos [1], osciladores [2] o controles combinatorios de expresión [3]. En cuanto al sondeo global, existen arreglos de oligonucleótidos con los cuales se puede conocer qué genes –de todo el organismo– estan expresados/inhibidos en distintas condiciones [4, 5, 6]. También existen métodos para sondear *in vivo*, la afinidad de los Factores de Transcripción (FT) al ADN, creando así redes de interacción FT-ADN [7]. Las técnicas anteriores se realizan a nivel de la colonia pero ya se desarrollan métodos que sondean, en células individuales, los niveles de proteína presente [8].

Los avances experimentales y la integración del conocimiento acumulado a lo largo de decadas de biología molecular han revelado grandes redes de interacción de los distintos componentes de la célula: redes metabólicas [9, 10], redes de regulación [11, 12, 13, 14], redes de señalización [15, 16], etc. Esto ha otorgado un nuevo panorama que obliga a enfatizar el hecho de un todo integrado, complementando la visión clásica que centra la importancia en las partes constituyentes. El nuevo énfasis y la cantidad de datos generados con las nuevas técnicas han propiciado el surgimiento de grandes bases de datos, formas de representarlos/estructurarlos y modelos que los expliquen.

De las redes anteriores, estamos particularmente interesados en las de regulación genética. Éstas son las encargadas del control, la memoria y la respuesta a largo plazo de la célula: coordinan la expresión de las proteínas

necesarias para sobrevivir en distintas condiciones y determinan la diferenciación celular, la apoptosis y la división de acuerdo a las distintas señales externas/internas. Hay dos enfoques detrás de los modelos dinámicos de estas redes: uno interesado en reproducir los detalles de la regulación como muestra de que se comprenden bien todos los mecanismos [17, 18] y otro interesado en saber cómo surge un orden superior capaz de coordinar a la célula [19, 20]. El primero se ve obligado a estudiar redes chicas ya que el detalle –aún con una cantidad modesta de genes (variables)– complica mucho el modelo haciendolo difícil de resolver o calcular. En el segundo, el tamaño del modelo es menos restrictivo ya que el detalle es secundario, sin embargo esto provoca que algunos resultados, respecto al modelo detallado, sean problemáticos de interpretar.

En el presente trabajo nos aproximamos a las redes de regulación según el segundo enfoque. La evidencia experimental muestra, de un forma cada vez más contundente, que la diferenciación celular y la estabilidad de distintos fenotipos –como el cáncer– está dada por una red regulatoria muy grande que subyace a estos procesos [21, 22]. Por lo tanto, entender a las redes de regulación desde un punto de vista global para saber cómo surge una organización de orden superior es la antesala al control de muchos mecanismos celulares complejos.

Las propiedades emergentes no se estudian exclusivamente desde la regulación genética. Sin embargo, el padre del modelo canónico de sistema complejo, Stuart Kauffman [23], obtuvo su inspiración inicial al querer entender cómo es que cientos o miles de genes pueden dar lugar a tipos celulares estables. Esto originó la propuesta de estudiar una red de variables binarias interconectadas pudiendo unas influenciar a las otras: una red genética donde los genes pueden estar expresados o inhibidos [24]. En la actualidad este modelo es ejemplo de cómo un sistema puede autoorganizarse, cómo una parte del mismo puede afectar al todo de forma no lineal además de resaltar que no hay que fijarse tanto en las propiedades de las partes sino en las propiedades del sistema como un todo.

Usando el modelo de Kauffman, mostramos que las redes reales de regulación operan alrededor de la fase crítica, Cap. 3. Operar en esta fase otorga la clara ventaja evolutiva de adaptación y, simultáneamente, estabilidad. Para mostrar lo anterior, implementamos modelos Booleanos de las redes reales y caracterizamos su dinámica global. Realizamos la implementación Booleana usando las redes que se conocen de tres microorganismos (*Escherichia coli*,

Bacillus subtilis y *Saccharomyces cerevisiae*) y cientos de microarreglos. A partir de estos datos, calculamos el valor más probable de la probabilidad de expresión genética, p, con inferencia paramétrica Bayesiana. El valor de p indica qué tan probable es que un gen se exprese estando sus reguladores en una configuración de expresión/inhibición particular. Implementamos el modelo dinámico de las redes, proponiendo conjuntos de funciones aleatorias acordes con el valor de p encontrado. La caracterización dinámica se lleva a cabo al medir la distancia de Hamming entre configuraciones similares cuando éstas evolucionan en la red. Hay tres casos: cuando la dinámica es caótica la distancia entre configuraciones aumenta, cuando es ordenada disminuye. Finalmente, cuando es crítica se conserva. El último caso se encuentra en la transición entre orden y caos. Nuestros resultados indican que las redes reales se encuentran alrededor de la transición. Lo mismo sucede en la red de regulación del desarrollo floral de la planta *Arabidopsis thaliana* y en la red de regulación que controla la formación de los segmentos polares en *Drosophila melanogaster*. El modelo Booleano de ambas ya ha sido implementado en otra parte [25, 26, 27]. La ubicuidad de la criticalidad en organismos tan dispares hace pensar que esta importante característica dinámica es genérica.

Por otra parte –suponiendo que el modelo de Kauffman captura las propiedades esenciales de la regulación–, mostramos que las redes, bajo la perturbación de duplicación seguida de divergencia genética, son robustas y evolucionables, Cap. 4. La perturbación consiste en añadir un nuevo gen a la red. Esto provoca que el paisaje de atractores se transforme. En particular, los atractores se fusionan, dividen, ganan configuraciones, las pierden, etc. pero también pueden quedar idénticos, sucediendo esto con mayor probabilidad en redes ordenadas y críticas. Además de los cambios anteriores, también pueden surgir nuevos atractores o desaparecer completamente. La conservación de los atractores inicialmente codificados por la red se identifica con el concepto de robustez; el surgimiento de nuevos con el de evolucionabilidad. Tanto la robustez como la evolucionabilidad son máximas en redes críticas. Creemos que estos resultados muestran que la estabilización de configuraciones transientes en nuevos atractores –sin que esto implique la pérdida de los ya codificados– es una propiedad esencial de las redes de regulación biológicas.

Capítulo 2

Modelo de Kauffman

Hace ya más de 40 años, Jacob y Monod mostraron que una red de regulación con sólo dos genes que se reprimen mutuamente, A y B, puede hacer las veces de un *switch* que mantiene dos tipos celulares estables: A expresado y B inhibido o viceversa [28]. Stuart Kauffman llevó al extremo esta bella idea explorando las propiedades de redes mayores y prediciendo la diferenciación de muchos más tipos celulares [24].

El modelo consiste de un conjunto de N variables binarias σ_1, σ_2, $\sigma_3, \ldots$, σ_N cada una representando a un gen expresado (1) o inhibido (0). Cada gen σ_n es regulado por k_n genes del mismo conjunto. A su vez, el gen σ_n puede regular a l_n genes; esto forma una red dirigida. Llamamos conectividad de entrada del gen σ_n al número de reguladores k_n y conectividad de salida al número de regulados l_n.

En una red dirigida, la topología de la conectividad de entrada y de salida puede ser distinta. En particular, si $P_e(k)$ denota la distribución de la conectividad de entrada y $P_s(l)$ denota la de salida, las redes reales tienen aproximadamente la siguiente topología (ver Apéndice B): P_e es una Poissoniana, $P_e(k) = e^{-K}K^k/k!$, con K el número de reguladores promedio por gen; P_s es una distribución Libre de Escalas, $P_s(l) = Cl^{-\gamma}$, con C una constante de normalización y γ un exponente cuyo valor depende de cada red particular. La primera distribución indica que existe una cantidad típica de reguladores por gen –en las redes reales, dos–, la segunda indica que no es improbable encontrar genes que regulen a un número arbitrario de otros genes, en particular no es raro encontrar genes que regulen a gran parte de la red. Es importante señalar que en el modelo de Kauffman original, cada gen

(de un conjunto de N) es regulado exactamente por K reguladores escogidos al azar. Esto provoca que la conectividad de salida sea Poissoniana. A este modelo se le conoce como *Modelo NK*.

La dinámica de la red se implementa al actualizar de manera sincrónica a todos los genes de la red de acuerdo a,

$$\sigma_n(t+1) = f_n(\sigma_{n_1}(t), \sigma_{n_2}(t), \ldots, \sigma_{n_{k_n}}(t)) \tag{2.1}$$

con f_n la función Booleana del gen σ_n que dicta cómo éste reacciona (expresión o inhibición) ante el estado combinado de sus reguladores en el paso anterior. Para cada una de las 2^{k_n} configuraciones de los reguladores de σ_n, $f = 1$ con probabilidad p y $f = 0$ con probabilidad $1 - p$. A p se le llama la probabilidad de expresión genética. Además, a f se le suele escribir como una tabla con las configuraciones posibles de los reguladores a la izquierda y el valor que toma la función a la derecha. Para cada gen σ_n se escoge una función f de todas las posibles del *ensemble*. Una vez hecha la elección de reguladores y función Booleana para cada gen, estos no cambian a lo largo de la dinámica. En la literatura, a esta situación se le conoce como *modelo no atemperado*[1].

A una configuración dada de los reguladores de un gen σ seguida del valor que toma la función Booleana, f, se le denomina una *frase de regulación*. Por ejemplo, si el gen σ_A es regulado por los genes σ_B y σ_C y la función Booleana f_A es tal que $f(0,0) = 0$, $f(1,0) = 1$, $f(0,1) = 1$, $f(1,1) = 1$, tenemos las siguientes frases de regulación: $00 \to 0$, $10 \to 1$, $01 \to 1$ y $11 \to 1$. Notemos que podemos estimar con qué valor de la probabilidad de expresión genética, p, fueron construidas las funciones Booleanas al conocer un conjunto de frases regulatorias y evaluar que fracción de éstas es activadora.

Podemos ver que, de una red regulatoria real al modelo propuesto hay varias simplificaciones y suposiciones. A continuación comentamos algunas.

- La regulación genética es un proceso muy detallado que involucra varios pasos entre la expresión regulada de un gen y la síntesis de su proteína producto: transcripción, procesamiento del ARN primario, traducción, regulación postranscripcional, procesamiento de la proteína para estabilizar el plegamiento, degredación de la proteína y de los

[1] *Quenched model* en inglés. Este modelo es contrario al modelo atemperado, *annealed model*, en el que las funciones o los reguladores pueden cambiar a lo largo de la dinámica.

mensajeros, etc. En el modelo se obvia toda esta cadena causal y se aproxima por una causa única –la expresión de un gen– cuyo efecto es la inhibición/activación de otro gen a través de la proteína producto, un Factor de Transcripción[2].

- La actualización sincrónica supone que en promedio la tasa de transcripción de todos los genes y la tasa de traducción de los transcritos al producto es la misma. Sin embargo, esto no se cumple en general ya que existen, por ejemplo, promotores que son más fuertes que otros, genes que se expresan a una cierta tasa sin algún factor de transcripción y a otra distinta con la presencia del mismo, variedad de longitudes en genes y mensajeros, etc.

- Cuando hablamos de expresión binaria entendemos que hay dos categorías: baja expresión y alta expresión. Baja expresión, $\sigma = 0$, indica que la concentración del ARNm de σ no es suficiente para que el producto de la traducción realice su función; alta expresión, $\sigma = 1$, indica lo contrario. Nuevamente, esta suposición no es del todo cierta ya que existen genes que se expresan de forma diferencial.

- No existe preferencia alguna al decidir la función Booleana del *ensemble* que se asigna a cada gen. Aunque se ha argumentado que cierto tipo de funciones deben ser más probables que otras [19], hasta el momento no existe evidencia experimental que muestre alguna sobrerepresentación de algún subconjunto del *ensemble* de funciones posibles.

- Considerar que ha transcurrido alguna unidad arbitraria de tiempo entre configuraciones consecutivas en la dinámica es problemático. Es mejor pensar que cuando se pasa a la siguiente configuración es porque requerimientos previos se han satisfecho en la configuración anterior. Para aclarar esto, pensemos en la expresión cíclica de proteínas que controlan el cíclo celular. Supongamos que son tres A, B y C y que cada proteína provoca la síntesis de la siguiente. Dependiendo de los nutrientes en el medio podemos afectar la duración del cíclo acelerando o desacelerando el tiempo de expresión entre A, B y C. Esto nos imposibilita definir una unidad de tiempo entre la expresión de las

[2]No haremos uso extenso de los términos Factor de Transcripción (FT) y Gen Objetivo (GO) ya que, si bien se sobreentiende que un FT puede a su vez ser un GO, un GO no puede ser un FT. Usar estos términos asimétricos podría causar confusión. Nos apegamos –toda vez que sea posible– a gen regulado y gen regulador en el entendido de que un gen regulado puede a su vez regular y viceversa.

distintas proteínas que sirva para todos los medios. En éste caso –con generalización inmediata a una red cualquiera– el modelo de Kauffman es incapaz de capturar el hecho de la variación en el tiempo de expresión. Sin embargo, lo que si captura es que A precede a B, B a C, C a A, etc. sin importar los lapsos de tiempo entre las distintas expresiones.

Nuestra creencia en que el modelo de Kauffman captura características esenciales de la regulación a pesar de las simplificaciones arriba expuestas se basa en hechos experimentales y fenomenológicos del propio modelo que exponemos a continuación.

Denotamos por Σ_t a una configuración de la red en el paso t de la dinámica,

$$\Sigma_t = (\sigma_1(t), \sigma_2(t), \ldots, \sigma_N(t)).$$

Inicializando a la red en alguna configuración Σ_0, el sistema sigue una trayectoria

$$\Sigma_0 \to \Sigma_1 \to \Sigma_2 \to \ldots \to \Sigma_t \to \ldots$$

determinada por la dinámica de la red, Ec. 2.1. Esta dinámica estructura el espacio de configuraciones genéticas dividiéndolo en conjuntos disjuntos: las configuraciones de cada conjunto tienen la propiedad de que, bajo la dinámica, van a otra configuración del mismo conjunto. Además, en cada conjunto existe un subconjunto especial de configuraciones donde, comenzando en una configuración arbitraria del subconjunto, es posible llegar a cualquier otro del mismo; estos subconjuntos son los *atractores*. No todas las configuraciones de cada conjunto pertenecen al atractor pero comenzando en cualquiera de ellas se llega a él en un número finito de pasos; estas *configuraciones transientes* forman la *cuenca de atracción*. Una representación gráfica de la dinámica global de una red Booleana se logra al conectar dirigidamente a todas las configuraciones de un mismo conjunto si, bajo la dinámica, una va a dar a la otra [29, 30]. Esto es la representación del *paisaje de atractores* (ver Fig. 2.1), abstracción análoga a los paisajes epigenéticos de Waddington [31].

La metáfora del *paisaje* sintetiza varios hechos de la diferenciación celular: la división del espacio de configuraciones en un número discreto de cuencas de atracción se corresponde con la observación de que el conjunto de tipos celulares de un organismo es discreto; el tamaño de la cuenca de atracción se corresponde con la estabilidad de los tipos celulares y –la corresponden-

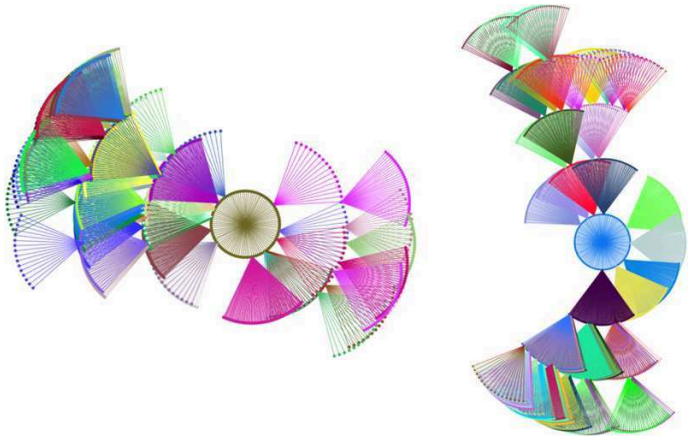

Figura 2.1: Paisaje de atractores para una red con 15 genes en la región crítica. En este caso, la red codifica dos cuencas. Cada punto representa una configuración de la red y uno está conectado a otro si bajo la dinámica el primero va a dar al segundo. Esto establece un flujo que converge en el centro de cada cuenca (atractor). Todos las configuraciones que tienen al mismo sucesor son del mismo color.

cia más atractiva de todas– los atractores son los tipos celulares estables[3]. Observemos que las constricciones dinámicas a las que está sujeta la red Booleana estructuran gratuitamente el espacio de configuraciones formando el paisaje de atractores –esto es parte de la noción de orden biológico que Kauffman ha legado y que es complementaria a la evolución por selección natural.

Existen hoy distintas pruebas experimentales que sustentan la visión de los tipos celulares como atractores de una red de regulación. Por ejemplo, existe un modelo Booleano de red regulatoria que sintetiza mucho del conocimiento genético que se tiene sobre el desarrollo floral de la planta modelo *Arabidopsis thaliana* [26, 27]. Esta síntesis esta implícita en la forma específica de las funciones Booleanas y en el conjunto de regulaciones que forman la red. Todos los atractores de la red Booleana corresponden a los patrones

[3]Esta correspondencia sólo tiene sentido biológico cuando la longitud de los atractores (número de configuraciones que conforman al atractor) es "pequeña". Si se aceptaran atractores arbitrariamente largos se podría objetar que ninguna célula acompletaría siquiera un periodo del atractor, perdiéndose, con esto, el sentido de los atractores como tipos celulares.

de expresión de distintos tipos celulares presentes en la flor. Más aún, si se simulan deleciones al inhibir la expresión de algún gen en la red, los atractores –bajo esta perturbación– reproducen los patrones de expresión mutantes. Algo similar sucede con la red de regulación que controla la formación de los segmentos polares en *Drosophila melanogaster*: los atractores de un modelo de red Booleano de 60 genes construido con todas las interacciones hasta ese momento conocidas reproducen los patrones de expresión que se observan a lo largo del embrión [25]. Nuevamente, mutaciones en la red que simulan mutaciones reales, originan nuevos atractores que se corresponden con los patrones mutantes de expresión. Los dos trabajos anteriores muestran la validez del modelo de Kauffman. Por otra parte, al suponer que el modelo es válido, se abre el camino para nuevos tipos de experimentos. Huang, dando por hecho que los tipos celulares son atractores, diseña experimentos que tratan de controlar y entender la diferenciación celular desde este punto de vista [32, 33, 34]. Los resultados obtenidos por los trabajos anteriores obligan a reconsiderar la importancia de todo el detalle bioquimico dado por las constantes de reacción y los pasos intermedios en la regulación para obtener respuestas relevantes [35].

Capítulo 3

Criticalidad

Existe evidencia de que muchos sistemas dinámicos complejos son críticos, es decir operan cerca de la transición entre dos regímenes dinámicamente distintos [36]. Avalanchas [37], fenómenos atmosféricos [38, 39], mercados financieros [40, 41], terremotos [42, 43], materia granular [44] y el cerebro [45, 46, 47] son algunos ejemplos típicos. Los sistemas críticos exhiben propiedades remarcables que serían difíciles de explicar sin asumir criticalidad. Por ejemplo, pueden integrar, procesar y transmitir información de forma rápida y robusta [48]. También pueden detectar y responder a estímulos externos cuya intensidad abarca varios ordenes de magnitud, como el cerebro [46]. Estas propiedades son principalmente consecuencia de las correlaciones de largo alcance que emergen cerca de la transición de fase produciendo comportamientos colectivos y respuestas coordinadas del sistema entero. La criticalidad confiere al sistema la habilidad de adaptarse y de responder colectivamente a un ambiente en permenante cambio.

En el caso de los organismos vivos, la experiencia nos muestra que estos son *estables* ante perturbaciones azarosas del exterior siendo al mismo tiempo *adaptables* a distintas condiciones ambientales. A nivel genómico, podemos encontrar la fuente de la adaptación y la estabilidad en las redes de interacción genética. La red debe, como un todo, integrar distintas señales externas para decidir una respuesta adecuada a nivel del organismo [49, 50, 21, 51, 52]. Esto significa que algunas señales deben disparar eventos especiales (adaptación) mientras que otras deben ser ignoradas (estabilidad). Para ejemplificar lo anterior, imaginemos a la bacteria *Escherichia coli* adaptada a un medio contaminado por algunas moléculas de glucosa jugando el papel de señal espuria. Si la red de regulación de *E. coli* fuera demasiado sensible a per-

turbaciones externas, cualquier señal desencadenaría vastos cambios en el patrón de expresión genético. En el caso de las pocas moléculas de glucosa, las proteínas necesarias para metabolizar la nueva fuente de energía serían producidas. Esto conllevaría un desperdicio energético no recuperable al procesar la glucosa. Por el contrario, si la red genética de *E. coli* ignorara cualquier cambio en la concentración de glucosa, la bacteria no aprovecharía la fuente de energía aún estando disponible.

El último ejemplo nos muestra *dos regímenes dinámicos* de la red regulatoría incompatibles con la experiencia que tenemos de los seres vivos: uno *caótico* demasiado sensible a cambios y otro *ordenado* difícil de perturbar. En realidad los organismos pueden cambiar e ignorar el ruido ambiental dentro de ciertos límites. Observaciones de éste tipo llevaron a Kauffman [53, 19] a formular la hipótesis de que la dinámica de los organismos vivos es *crítica*, *i.e.* se encuentra alrededor de la transición de fase entre un comportamiento ordenado y otro caótico.

En este capítulo nos enfocaremos a caracterizar la dinámica de las redes reales encontrando la primera evidencia experimental directa de su cercanía al estado crítico. La metodología es la siguiente: tomamos como referencia la dinámica del *ensemble* de redes Booleanas aleatorias, obtenemos el equivalente Booleano de redes reales y comparamos su dinámica con la referencia obtenida del *ensemble*. La referencia dinámica de redes Booleanas aleatorias se obtiene gracias a un cálculo de Derrida [54] generalizado por Aldana [55] que aquí reproducimos. La obtención del equivalente Booleano de las redes reales ocupa gran parte del capítulo. El mayor problema –resuelto usando inferencia Bayesiana y experimentos de microarreglo– es estimar la probabilidad de expresión genética p; valor que se usa para construir las funciones Booleanas. Finalmente, usando los datos obtenidos para cinco organismos (*Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster*) mostramos que la dinámica de estos seres vivos está muy próxima a la transición de fase orden/caós.

3.1. El cálculo de Derrida

En la introducción del capítulo sugerimos que la dinámica de la red puede ser ordenada, crítica o caótica. Podemos ver esto siguiendo la evolución de dos configuraciones ligeramente diferentes en el modelo de Kauffman. La dinámica es ordenada si, a lo largo de su evolución, las configuraciones con-

vergen. Si por el contrario divergen, la dinámica es caótica.

Comenzamos[1] definiendo una medida de la distancia entre dos configuraciones $\Sigma_t = \{\sigma_1(t), \sigma_2(t), \ldots, \sigma_N(t)\}$ y $\Sigma'_t = \{\sigma'_1(t), \sigma'_2(t), \ldots, \sigma'_N(t)\}$, con σ_i el estado del gen i en la red. La medida es la distancia de Hamming normalizada $D(t)$ entre las dos configuraciones Σ_t y Σ'_t,

$$D(t) = \frac{1}{N} \sum_{i=1}^{N} |\sigma_i(t) - \sigma'_i(t)|. \tag{3.1}$$

La distancia de Hamming normalizad es la fracción de genes distintos entre las dos configuraciones en el paso t de la dinámica. Si las configuraciones Σ y Σ' fueran aleatorias tendríamos $D(t) \approx 1/2$. En cambio, si fueran casi idénticas, $D(t) \approx 0$. Podemos considerar al límite estacionario de la distancia de Hamming, $D^* = \lim_{t\to\infty} D(t)$, como un parámetro de orden que caracteriza a la dinámica de la red. Si $D^* = 0$, el sistema es insensible a perturbaciones iniciales y diferencias entre configuraciones similares desaparecen. En este caso la dinámica de la red es ordenada. Cuando $D^* \neq 0$, diferencias en las configuraciones iniciales persisten alcanzando a una fracción no nula de todo el sistema; la dinámica de la red es caótica.

Podemos establecer la transición entre la respuesta ordenada y caótica encontrando la ecuación que dé la evolución de la distancia de Hamming. Sea $A(t)$ el conjunto de genes que son idénticos en las configuraciones $\Sigma(t)$ y $\Sigma'(t)$ y $B(t)$ el conjunto de genes distintos. Para continuar el cálculo, conviene definir el traslape x como $x = 1 - D$, es decir, la fracción de genes idénticos en ambas configuraciones. El traslape x es el número de genes en A dividido por el total de genes N. Existen dos posibilidades para los k_i reguladores de un elemento arbitrario σ_i:

- Los k_i reguladores de $\sigma_i(t)$ están en $A(t)$. Esto ocurre con probabilidad $x^{k_i}(t)$ e indica que el estado de los reguladores de σ_i y σ'_i es idéntico por lo que el valor de la función f_i también lo es concluyendo que $\sigma_i(t+1) = \sigma'_i(t+1)$.
- Por lo menos uno de los k_i reguladores de $\sigma_i(t)$ está en $B(t)$. Esto ocurre con probabilidad $1 - x^{k_i}(t)$. Los genes $\sigma_i(t+1)$ y $\sigma'_i(t+1)$ son idénticos sólo si la función f_i tiene el mismo valor para argumentos distintos. De la definición de las funciones Booleanas[2] sabemos

[1] El cálculo que aparece a continuación puede encontrarse en Aldana [55].
[2] Ver Cap. 2.

que, sin importar el valor de los reguladores, $f_i(\sigma_{i_1}(t), \ldots, \sigma_{i_{k_i}}(t)) = f_i(\sigma'_{i_1}(t), \ldots, \sigma'_{i_{k_i}}(t))$ con probabilidad $p^2 + (1-p)^2$.

Por lo tanto, la probabilidad $x(t+1)$ de que $\sigma_i(t+1) = \sigma'_i(t+1)$ es

$$x(t+1) = \sum_{k_i=1}^{\infty} \{x^{k_i}(t) + [1 - x^{k_i}(t)][p^2 + (1-p)^2]\} P(k_i), \tag{3.2}$$

con $P(k_i)$ la probabilidad de que el gen σ_i sea regulado por k_i genes. Reacomodando términos y usando el hecho de que $\sum_{k_i=1}^{\infty} P(k_i) = 1$ encontramos que el traslape obedece la ecuación

$$x(t+1) = 1 - 2p(1-p)(1 - \sum_{k=1}^{\infty} x^k(t) P(k)). \tag{3.3}$$

Sustituyendo $D = 1 - x$, llegamos a,

$$D(t+1) = 2p(1-p)(1 - \sum_{k=1}^{\infty} (1 - D(t))^k P(k)). \tag{3.4}$$

Hemos encontrado que la distancia de Hamming está determinada por la ecuación $D(t+1) = M(D(t))$ con M el mapeo,

$$M(D) = 2p(1-p)(1 - \sum_{k=1}^{\infty} (1 - D)^k P(k)), \tag{3.5}$$

al que se denomina comunmente como *Mapeo de Derrida* [55].

La condición para encontrar el punto fijo es $D^* = M(D^*)$. Notemos que M es una función monótonamente creciente con $M(1) = 2p(1-p)$ y $M(0) = 0$. Por la última ecuación podemos darnos cuenta que $D^* = 0$ siempre es un punto fijo del mapeo M. Sin embargo, la estabilidad de este punto fijo puede cambiar dependiendo de los parámetros de la red. Es fácil ver que si $\lim_{D \to 0^+} dM/dx < 1$ el punto $D^* = 0$ es el único punto fijo y es estable. Esta es la fase ordenada. Cuando $\lim_{D \to 0^+} dM/dx > 1$, el punto $D^* = 0$ se vuelve inestable y aparece otro punto fijo $D^* > 0$ estable. Esta es la fase caótica, ver Fig. 3.2. La fase crítica se encuentra en la transición entre la fase caótica y la ordenada, $\lim_{D \to 0^+} dM/dx = 1$, en la cual el mapeo M se aproxima asintóticamente a la identidad y es tangente a ésta en $D^* = 0$. Derivando la Ec. 3.5 llegamos a una ecuación explícita para conocer el estado dinámico de la red,

$$\lim_{D \to 0^+} \frac{dM}{dD} = 2p(1-p)K, \tag{3.6}$$

con K igual al promedio de la distribución $P(k)$. A esta ecuación se le conoce como la *sensibilidad*, $S = 2p(1-p)K$, de la red.

Podemos entender de forma intuitiva el significado del cálculo de Derrida al comparar los patrones de expresión en microarreglos de un organismo silvestre y su mutante, ver Fig. 3.1. En este caso, la mutación (*e. g.* algunas deleciones) es una pequeña perturbación a la red genética, ver Fig. 3.1 *a*. Por ejemplo, si el organismo tiene 1,000 genes, 10 deleciones representan una diferencia inicial del 1 % entre el patrón silvestre y el mutado, *i. e.* $D(t = 0)$ =0.01. Cuando el organismo es ordenado y han transcurrido algunos pasos en la dinámica, las deleciones, en promedio, no causan ningún cambio en el patrón de expresión mutante siendo éste y el silvestre idénticos, $D^* = 0$, ver Fig. 3.1 *b*. Por el contrario, cuando el organismo es caótico, las deleciones de unos cuantos genes (después de unos pasos en la dinámica) provocan vastos cambios entre el patrón de expresión mutante y el silvestre $D^* \neq 0$, ver Fig. 3.1 *c*. En el caso ordenado tenemos que la red reacciona siempre de la misma forma sin importar cual sea la perturbación inicial por lo que es incapaz (la red) de detectar cambio alguno (sensibilidad $S < 1$). Por el contrario, en el caso caótico cualquier cambio por pequeño que sea es amplificado hasta perturbar a toda la red (sensibilidad $S > 1$). Sólo cuando el organismo es crítico las deleciones provocan cambios que se mantienen del mismo tamaño que la perturbación inicial, ver Fig. 3.1 *d*. Gracias a esta característica las señales pueden detectarse y transmitirse provocando cambios controlados en el organismo (sensibilidad $S = 1$).

La Ec. 3.6 determina el estado dinámico de la red con sólo conocer dos parámetros globales: el promedio de reguladores por gen K y la probabilidad de expresión genética p. Poder caracterizar la dinámica de una red con sólo dos parámetros es posible porque en el cálculo asumimos que todos los genes son independientes y estadísticamente equivalentes. Sin embargo, esta suposición es incorrecta en redes reales por la existencia de reguldores globales. La ocurrencia de estos reguladores podría correlacionar gran parte de la expresión de la red provocando la desaparición de la transición de fase. Para saber cómo afecta esta particularidad a la transición de fase predicha por el cálculo de Derrida-Aldana, hacemos simulaciones numéricas con redes construidas según la topología real, ver Apéndice B.

Para comparar la transición analítica con los resultados numéricos, primero sustituimos $P(k)$ en la Ec. 3.5 por una distribución de Poisson, $P(k) = e^{-K}K^k/k!$, que es la distribución de la conectividad de entrada de las redes

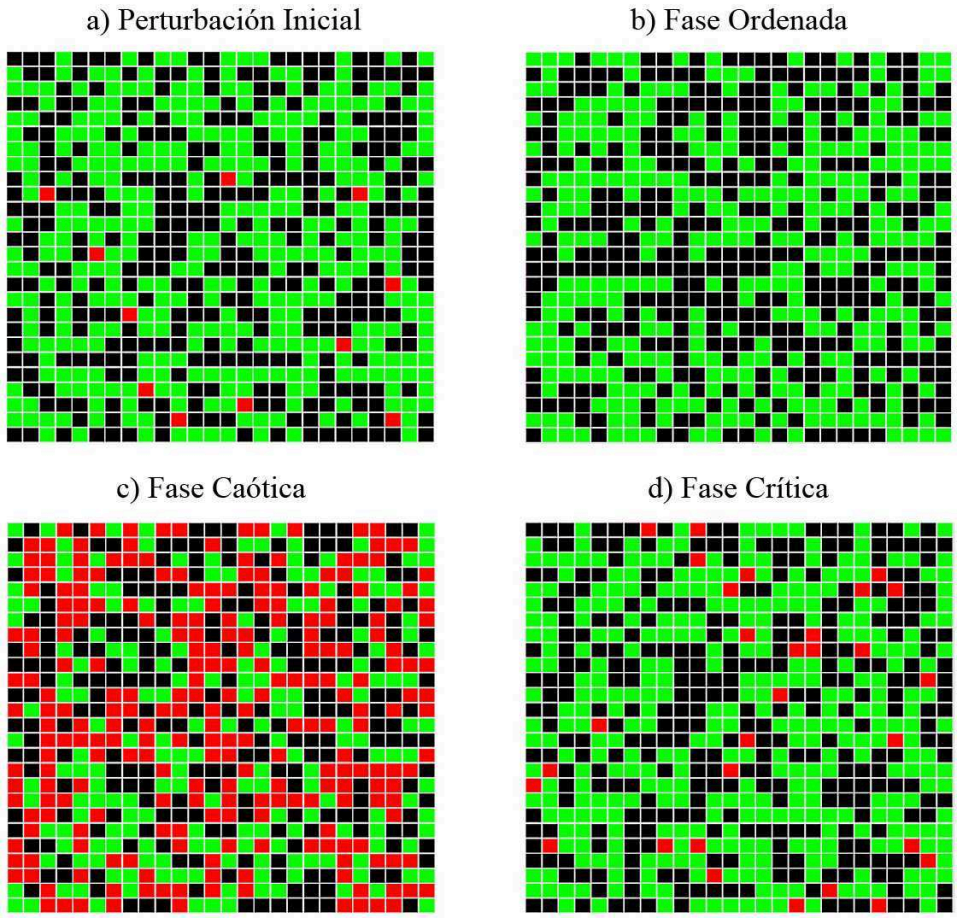

Figura 3.1: El cálculo de Derrida visto como una perturbación detectada a través de la comparación entre microarreglos de la expresión de un organismo silvestre y su mutante. Cada gen del organismo es un cuadro del microarreglo. Cuando éste es verde, el gen se expresa tanto en la cepa silvestre como en la mutante. Cuando es negro, no se expresa ni en la silvestre ni en la salvaje. Cuando es rojo, el gen se expresa en la cepa silvestre pero no en la mutante o viceversa. *a)* Microarreglo que muestra la perturbación inicial como una pequeña fracción de genes cuya expresión no coincide entre la cepa silvestre y la mutante. Después de algunos pasos en la dinámica suceden tres cosas dependiendo de la fase en la que se encuentre la red. *b)* Fase ordenada: la perturbación inicial desaparece y el patrón de ambas cepas coincide completamente. *c)* Fase caótica: la perturbación se ha propagado a toda la red y persiste durante toda la dinámica. *d)* Fase crítica: la perturbación persiste sin afectar a toda la red.

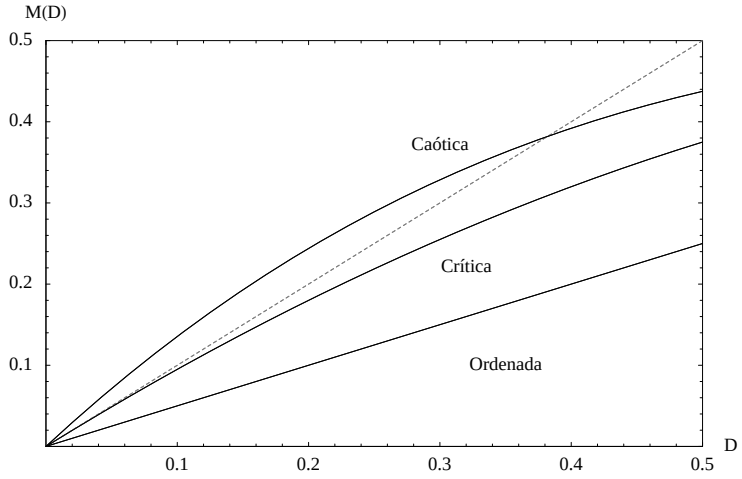

Figura 3.2: El mapeo de Derrida es monótonamente creciente. Notemos que $D^* = 0$ siempre es un punto fijo del mapeo. Sin embrago, en la curva caótica, la pendiente en $D^* = 0$ es mayor a uno por lo que el punto fijo es inestable. De las tres curvas, la caótica corta a la identidad en un segundo punto fijo $D^* > 0$ ahora estable (derivada menor a uno). Por otra parte, vemos que la curva crítica llega tangente a la identidad en $D^* = 0$.

reales con K reguladores en promedio. Tomando lo anterior en consideración, el punto fijo D^* de la Ec. 3.5 es,

$$D^* = 2p(1-p)(1-e^{-KD^*}) \tag{3.7}$$

La solución de la ecuación anterior nos da la distancia final entre dos configuraciones arbitrarias en el límite de iteraciones infinitas en la dinámica de la red. En el caso particular con $p = 1/2$ y $S = K/2$ (ver Ec. 3.6) llegamos a

$$D^* = \frac{1}{2}(1-e^{-2SD^*}). \tag{3.8}$$

Cuando la red es ordenada, el valor estacionario de la distancia D^* es cero: la perturbación desaparece. Cuando es caótica, el valor es distinto de cero: la perturbación inicial permanece en la red. Las redes críticas se encuentran en la interfaz de estos dos comportamientos.

En la Fig. 3.3 hemos graficado la Ec. 3.8 y la distancia D^* resultado de simulaciones para redes con topología Ley de Potencias en la conectividad de salida para tamaños de red que abarcan tres ordenes de magnitud[3]. A medida que el número de elementos que conforman la red aumenta, la curva resultado de la simulación va convergiendo al resultado de Derrida que es para el caso de redes infinitas. Concluimos que, aún no estando considerada la peculiaridad topológica de reguladores globales, el cálculo de Derrida aún es válido sirviendo de control para comparar la dinámica de las redes reales.

3.2. Estimación de p

En la sección anterior hemos vísto como, conociendo el mapeo dinámico M de la distancia de Hamming D, es posible determinar la dínamica de la red. En las redes reales, el mapeo M está definido por la topología particular de la red y por el valor de las funciones de Booleanas. Buena parte de la topología particular de algunas redes es conocida gracias a miles de experimentadores que han establecido regulaciones individuales y al trabajo de curadores que han organizado dichas regulaciones en grandes redes. Tal es el caso de las redes de *E. coli* [13], *B. subtilis* [14] y *S. cerevisiae* [11], por ejemplo. Sin embargo, las funciones Booleanas son desconocidas. Ésto nos obliga a generar funciones aleatorias acordes al valor de la probabilidad

[3]El cálculo de la distancia D^* implica

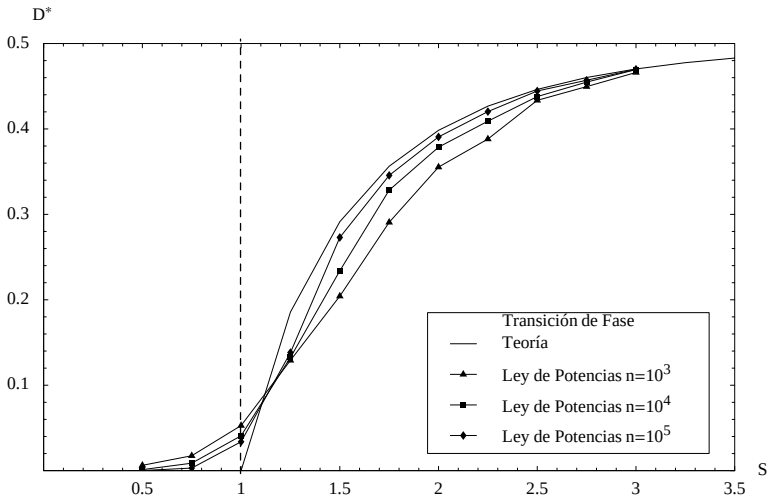

Figura 3.3: Aún en redes aleatorias con distribución Libre de Escalas en la conectividad de salida, la transición de fase encontrada por Derrida se conserva. Cuando la sensibilidad es menor a uno, la distancia estacionaria D^* es cero (orden). A partir de $S = 1$, D^* es positiva (caos). La transición orden-caos ocurre en $S = 1$; n indica el tamaño de la red. Cada punto es el resultado de promediar sobre 500 redes la distancia entre 100 pares de configuraciones que difieren en un 10 % inicialmente. La distancia D^* entre cada par de configuraciones es el promedio del valor de $D(t)$ en 500 pasos de la dinámica después de que ésta ha evolucionado 2,500 pasos (transiente).

de expresión genética p. El valor de p lo estimamos usando inferencia paramétrica Bayesiana alimentada por cientos de microarreglos.

Esta sección está dividida en tres partes. En la primera, mostramos cómo inferir frases de regulación, discutimos la conexión entre funciones de cambio y de estado, comentamos sobre la selección de los microarreglos, caracterizamos los ciclos dirigidos en las redes reales y mostramos que es imposible inferir todas la frases de regulación aún teniendo experimentos en todas las condiciones posibles. En la segunda parte, concretamos todo lo anterior en un algoritmo dando la descripción de cada uno de los pasos principales. Finalmente, en la tercera parte, estimamos el valor de p de tres microorganismos con los resultados obtenidos al aplicar el algoritmo de inferencia a conjuntos de microarreglos.

3.2.1. Inferencia de Funciones de Cambio

Para llevar a cabo la inferencia de las tablas de cambio de una red (definidas más adelante), podemos fragmentar a ésta –siempre y cuando no contenga ciclos dirigidos[4]– en varias subredes, cada una de ellas compuesta por un nodo y todos los nodos que pueden afectar directamente al primero [56, 57, 58]. En una red de regulación lo anterior se traduce en que podemos fragmantar a la red en subredes de reguladores-regulado; la subdivisión creará tantas subredes como genes regulados haya. Por este efecto, de ahora en adelante, sólo nos enfocaremos en una subred cualquiera.

Supongamos que sabemos que un gen A es regulado *sólo* por dos genes B y C y que no sabemos nada más sobre el circuito. Dádo el grafo anterior nos interesa saber cómo B y C regulan a A dado un conjunto de evidencias como información. Las evidencias serán experimentos de microarreglo y sólo podremos medir los *cambios* en los niveles de expresión genética. Estos cambios son de dos tipos: aumento o disminución del nivel de expresión. Como sólo hay dos tipos de cambio es fácil enlistar en una *tabla de cambio* las distintas formas en que el cambio combinado de B y C puede afectar a A y las formas en que A responde,

[4]En una red, un ciclo dirigido es un camino que lleva de un nodo A a si mismo pudiendo pasar (el camino) por nodos intermedios o por ningún nodo. En el caso de las redes de regulación, cuando un gen se autoregula tenemos un ciclo que se denomina *asa de retroalimentación*. Cuando el ciclo en la red contiene más de un gen, se le denomina *ciclo de regulación*.

B	C	A
$\downarrow$	$\downarrow$	-
$\uparrow$	$\downarrow$	-
$\downarrow$	$\uparrow$	-
$\uparrow$	$\uparrow$	-

Las columnas del lado izquierdo de la línea vertical enlistan todas las formas en las que B y C pueden cambiar; $\downarrow$ indica disminución en el nivel de expresión, $\uparrow$ aumento. Al lado derecho de la tabla indicamos con una línea recta horizontal que no sabemos cómo reacciona A ante cada cambio combinado de B y C.

Para conocer la respuesta de A usamos experimentos de microarreglo, Fig. 3.4. En esta misma figura, la tabla de cambio anterior se ve modificada y ahora en la columna de A contabilizamos el número de *evidencias* en el microarreglo que soportan a alguna de las dos posibilidades, $A\uparrow$ o $A\downarrow$, para cada cambio combinado de B y C. En la Fig. 3.4.*a* mostramos la *inicialización* de la tabla a la que llenamos con 1's para representar nuestro conocimiento *a priori* de la regulación[5]. Para cada combinación de B y C hemos contabilizado una evidencia tanto para el caso $A\downarrow$ como para el caso $A\uparrow$. Asignando una evidencia no damos preferencia *a priori* a un caso o al otro, lo que refleja ignorancia total en cómo A es regulado[6].

Inicializada la tabla comienza la contabilización de la ocurrencia de los distintos casos según aparecen en los microarreglos, Fig. 3.4.*b*. Las distintas ocurrencias se suman a las ocurrencias establecidas *a priori*. En el ejemplo, podemos ver que al final de la contabilización, *b*.4, hemos sumado una evidencia a la combinación $A\uparrow B\uparrow C\uparrow$ y a la combinación $A\downarrow B\downarrow C\downarrow$ y dos evidencias a la combinación $A\uparrow B\downarrow C\uparrow$. La tabla en este paso contiene nuestro conocimiento sobre la regulación *a posteriori*. A partir de ella decidimos el tipo de regulación que B y C ejercen sobre A en base al porcentaje de evidencias que indican una activación/inhibición en cada una de las

[5]En realidad, podríamos inicializar la tabla con cualquier valor positivo. La tabla no puede ser inicializada con 0's porque podría haber indefiniciones en cálculos posteriores (divisiones de cero entre cero) como se verá más adelante.

[6]Podría suceder que tuviéramos información regulatoria *a priori* que favoreciera a alguno de los dos cambios posibles de A. Por ejemplo, si en la literatura encontráramos cuatro referencias que reportan la respuesta de A ante cierto cambio combinatorio de B y C y en tres de éstas se reportara un aumento y en la restante una disminución, entonces nuestra información *a priori* podría reflejar este hecho al inicializar la tabla con tres evidencias asignadas a $A\uparrow$ y una a $A\downarrow$.

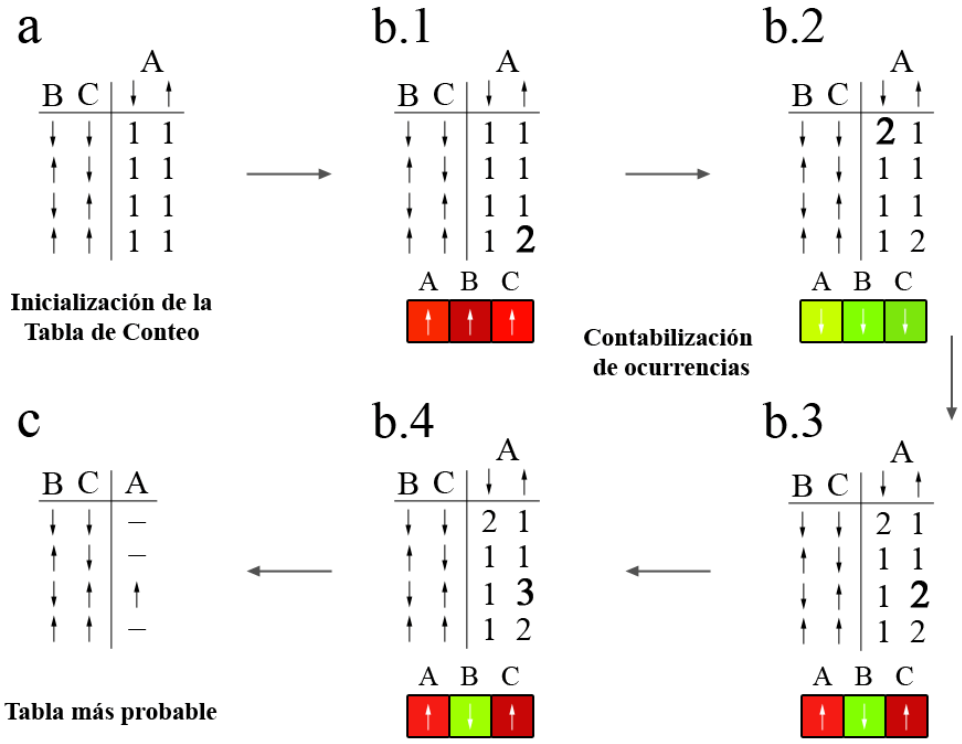

Figura 3.4: Ejemplo con tres genes (un regulado A y dos reguladores B y C) que muestra la esencia de la inferencia de las tablas de cambio. *a.* Inicialización de la tabla de conteo con evidencias (cambios) *a priori*. *b.* Contabilización de los cambios correlacionados detectados en cuatro experimentos de microarreglo. Cada microarreglo está representado esquemáticamente abajo de cada etapa de la contabilización. Los colores representan: rojo (↑) un aumento en la expresión, verde (↓) una disminución en la expresión. *c.* Finalizada la contabilización, se decide, en base a la evidencia que apoya inhibición/activación, los cambios más probables llegando a la tabla de cambio.

frases. Notemos que, cuando se inicializa la tabla, hay dos evidencias para cada combinación de B y C que apoyan a cada uno de los cambios posibles de A por lo que los porcentajes de evidencia para la activación o inhibición son idénticos, 50 %, y no podemos decidir. Contabilizadas las distintas ocurrencias, el porcentaje de evidencias que indican activación de A para cada combinación de B y C son: $A \uparrow B \downarrow C \downarrow = 33\,\%$, $A \uparrow B \uparrow C \downarrow = 50\,\%$, $A \uparrow B \downarrow C \uparrow = 75\,\%$, $A \uparrow B \uparrow C \uparrow = 66\,\%$. El porcentaje de evidencias que indican inhibición es simplemente uno menos el porcentaje de evidencias que indican activación.

Para decidir el tipo de regulación de A debemos usar un límite artbitrario. Para este ejemplo si un 70 % o más de las evidencias indican que A aumenta su nivel de expresión entonces diremos que hay suficiente evidencia para creer que A es activado; si un 30 % o menos de las evidencias indican que A aumenta su nivel de expresión entonces diremos que A es inhibido. Si el porcentaje de activaciones no cae en alguno de los rangos considerados, los datos no apoyan un tipo u otro de regulación. Considerando esto último llegamos a la tabla de cambio más probable según las evidencias experimentales, Fig. 3.4.*c*. Notemos que para tres combinaciónes de B y C y un umbral del 70 % no podemos determinar el tipo de regulación sobre A con la certeza deseada; sólo para la combinación de $B \downarrow C \uparrow$ tenemos que $A \uparrow$. En este último caso decimos que hemos determinado una *frase de regulación*.

Hasta el momento, hemos hablado de cómo inferir frases de regulación usando cambios correlacionados en experimentos de microarreglo. Sin embargo, el modelo de Kauffman no considera tablas de cambio (aumento/disminución) sino de estado (presente/ausente), mismas que necesitamos para estimar el valor de p. Para hacer la conexión que nos lleve de las tablas de cambio a las de estado notemos que los niveles de expresión absolutos tienen límites impuesto por la biológia de cada organismo. El límite inferior del nivel de expresión de un gen es aquel nivel mínimo por debajo del cual el organismo muere o por debajo del cual la expresión del gen no tiene efecto. Análogamente, el límite superior es el nivel máximo fisiologicamente posible. A estos límites los identificamos como el estado cero o uno. Notemos que el carácter de cada frase de regulación (activación o inhibición) no cambia al extremizar los cambios detectados hacia su expresión máxima o mínima para cada uno de los genes que participan. El carácter sólo depende de cómo interactuan los factores de transcripción con la región reguladora del gen regulado y no de su cantidad. Por todo lo anterior, cuando más adelante calculemos el valor de p para distintos organismos usando las tablas de

cambio inferidas tendremos en mente que $\downarrow$ equivale a 0 y que $\uparrow$ equivale a 1.

La efectividad de la inferencia arriba ejemplificada depende de varias cosas, a saber: de la no existencia de causas ocultas, de la adecuada selección de los experimentos de microarreglo y de la validez del procedimiento en una red con ciclos. Aún asegurando que todo lo anterior se cumple o se evita según sea el caso, existen frases de regulación que no se pueden inferir.

No Existencia de Causas Ocultas

Esto quiere decir que los reguladores representan todas las causas distintas por las que puede cambiar su expresión el gen regulado y no existe ninguna otra. En general supondremos que ésto se cumple. No obstante, debemos tener presente que las redes reportadas son aún incompletas y que existen genes de los que no se conoce el conjunto completo de sus reguladores. Estos reguladores no reportados (causas ocultas) pueden provocar que las frases inferidas con los reguladores reportados arrojen resultados contradictorios o inesperados. Otra fuente de causas ocultas que no tiene que ver con la incompletez de la red de transcripción y que aquí no consideramos es la dependencia de la regulación con el contexto celular. Se sabe que existen Factores de Transcripción que cambian su rol de activador o inhibidor dependiendo de la condición en la que se encuentra el organismo [59]. Por lo tanto, sería necesario extender la red de regulación para, además de considerar a los reguladores como nodos en las redes, también considerar nodos que representen a los distintos contextos celulares.

Selección de los Microarreglos y su Ruido Intrínseco

En general, al escoger los experimentos de microarreglo se tiene que tratar de cumplir con que éstos exploren una amplia variedad de condiciones y no descartar la inclusión de experimentos replicados. Contar con un espectro amplio de experimentos en las distintas condiciones en las que un organismo puede sobrevivir aumenta la posibilidad de que haya conteos en distintas frases de regulación ya que se explora un área más grande del espacio de expresión. De igual importancia es considerar réplicas y no desechar experimentos de condiciones que pudieran ser redundantes. Esto tiene como consecuencia que nueva evidencia puede respaldar a la anterior haciendo a la inferencia más robusta ante los falsos positivos/negativos debidos al ruido intrínseco de los microarreglos [4, 5, 6, 60]. A causa de este mismo ruido, no siempre es posible extraer información para todos los genes en un sólo

experimento. En el ejemplo de inferencia, Fig. 3.4, hemos supuesto que podemos establecer sin ambiguedad, para cada gen individual, alguno de los dos cambios, aumento o disminución de la expresión. Sin embargo, debido al ruido intrínseco de en los microarreglos, las intensidades de cambio deben sobrepasar cierto umbral para tomarse como cambios relevantes. Todo lo anterior lo hemos considerado en la selección de los datos con los que trabajaremos más adelante.

Ciclos en las Redes Reales

En el ejemplo de inferencia hemos pedido explícitamente que la red a inferir no contenga ciclos de regulación en la red dirigida aunque sepamos que las redes reales los presentan. Esto es un requerimiento de la inferencia paramétrica en redes Bayesianas independentes del tiempo [56, 58]. Debemos mencionar que aún existiendo metodología para tratar los ciclos de regulación de forma adecuada, tenemos el inconveniente de que los datos necesarios de microarreglo son insuficientes o no existen. La metodología (que involucra redes dinámicas Bayesianas) requiere de series temporales de microarreglos en donde todos los miembros de la colonia hayan sincronizado su ciclo de división celular. Desafortunadamante, este tipo de experimento sólo es posible en colonias de *S. cerevisiae*[7] [61]. Otro inconveniente es que no existe un espectro amplio de condiciones exploradas, algo importante para el éxito de la inferencia como hemos visto.

Por lo anterior debemos justificar que ignorar los ciclos afecta en poco la inferencia. Adelantamos que esto se debe a que gran parte de la red tiene una estructura jerárquica que es *controlada* por un pequeño conjunto de nodos donde se encuentran *algunos pocos* ciclos. Kauffman ha especulado sobre esta estructura y la ha llamado red *medusa* [62]. En la Fig. 3.5 proponemos un procedimiento para revelar la *cabeza* de la red *medusa*. Primero eliminamos todos los nodos cuya conectividad de salida sea cero. Esto revela nuevos nodos con conectividad de salida cero que también eliminamos. Continuamos este proceso hasta que ya no se pueda eliminar ningún nodo. Todos los nodos en la red así obtenida influyen por lo menos a otro nodo aunque no todos reciben influencias, es decir, su conectividad de entrada

[7]Existen dos técnicas experimentales para lograr la sincronización de la colonia: selección por centrifugación-elutriación y el uso de mutantes cuyo ciclo es sensible a la temperatura. La primera genera una población homogénea de pequeñas células que comienzan su ciclo celular. La segunda detiene la división en una fase específica del ciclo celular (al variar la temperatura) a partir de la cual las células recomienzan una división sincrónica.

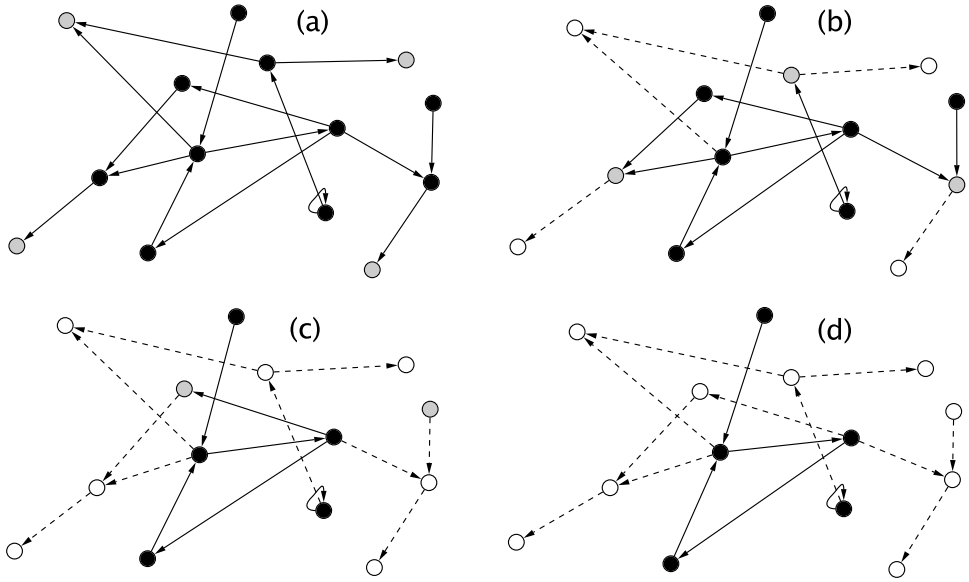

Figura 3.5: Procedimiento para obtener la cabeza de la red *medusa*. (a) Todos los nodos con conectividad de salida cero son marcados para su eliminación (nodos grises). (b) Los nodos grises en (a) son eliminados (nodos blancos), esto provoca la aparición de nuevos nodos con conectividad de salida cero que nuevamente marcamos en gris. (c) Repetimos los pasos en (b). (d) El proceso termina cuando ya no se pueden eliminar más nodos; la red restante es la cabeza de la red. Por la definición del proceso, si existe algún ciclo, entonces debe estar en la cabeza.

organismo	Genes en total	Genes en la *cabeza*	Genes en ciclos	Asas de retro-alimentación
B. subtilis	830	66	53	49
E. coli	1,328	103	89	87
S. cerevisiae	3,459	76	32	21

Cuadro 3.1: Genes de tres organismos que participan en la cabeza de la red. En la segunda columna enlistamos el número total de genes en las redes experimentales. En la tercera columna reportamos el número de genes en la cabeza de la red. Después mostramos el número de genes que participan en al menos un ciclo. La última columna muestra los genes que participan en un asa de retroalimentación (un gen que se autoregula). Notemos que casi todos los genes que participan en al menos un ciclo tienen un asa de retroalimentación.

puede ser cero. Por comodidad a estos nodos les llamaremos *nodos entrada*. Una propiedad interesante de la cabeza de la red es que contiene a todos los ciclos.

En el Cuadro 3.1 mostramos el número de nodos que participan en la cabeza y el número de asas de retroalimentación en las redes *incompletas* de *E. coli* [13], *S. cerevisiae* [11], *B. subtilis* [14]. Llama la atención que menos del 8 % de los genes en los tres organismos se encarguen del control de toda la red. Las asas de retrolimentación son los ciclos más pequeños, una busqueda de ciclós de orden mayor a uno da cómo resultado unos pocos circuitos que mostramos en la Fig. 3.6, en la Fig. 3.7 y en la Fig. 3.8.

Gracias al análisis topológico realizado, nos damos cuenta de que, hasta donde se conocen las distintas redes de regulación, su estructura es prácticamente jerárquica[8]: un gran conjunto de nodos salida o esclavos que son comandados por una red cabeza donde se encuentran pocos ciclos siendo el

[8]Este resultado topológico repercute positivamente en la extensión y los tiempos de cálculo computacional de la dinámica: Los nodos eliminados son esclavos que siguen las ordenes que se calculan de la dinámica de la cabeza y que no afectan a esta última. La dinámica de la cabeza, una vez establecido el contexto celular que se transduce a través de los nodos entrada, calcula la respuesta de la red. Desde la perspectiva de Kauffman, la dinámica de *toda* la red (en un contexto constante) viene dada por los atractores de la red cabeza. *La descomposición (esclavos, cabeza y ciclos en la cabeza) vuelve trivial el problema de calcular la dinámica de la red desde el punto de vista de recursos computacionales haciendo posible encontrar todo el conjunto de atractores compatibles con cierto contexto celular en menos de un minuto con una computadora típica de escritorio.*

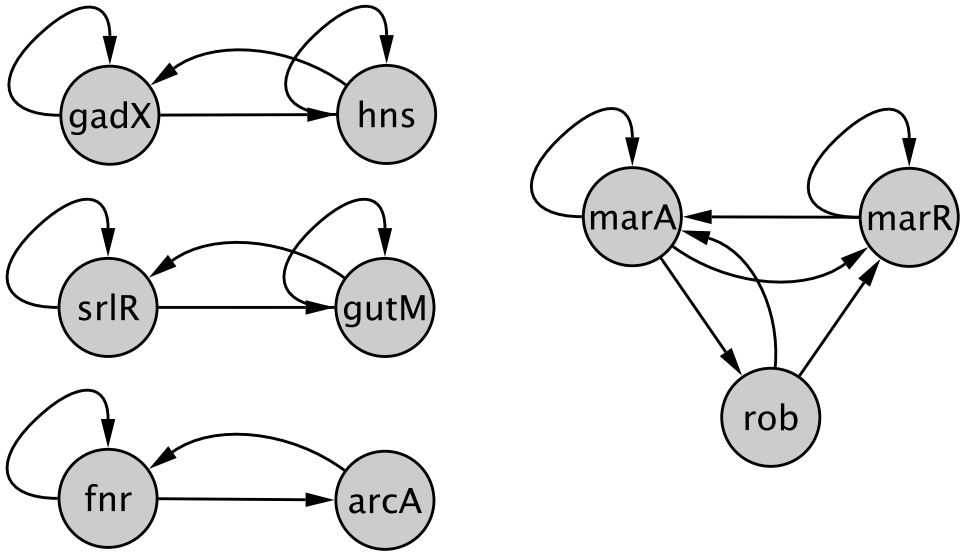

Figura 3.6: Circuitos que contienen ciclos dirigidos de orden mayor a uno en *E. coli*.

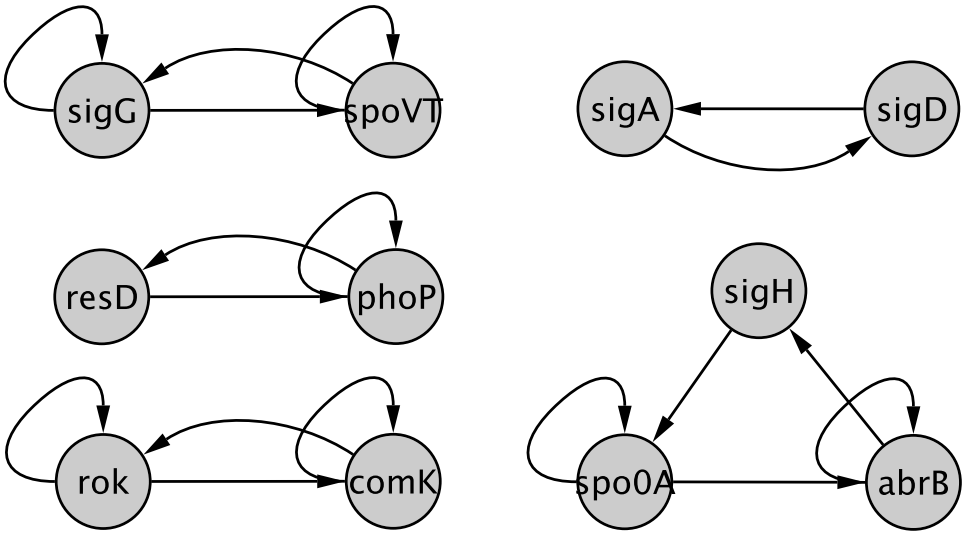

Figura 3.7: Circuitos que contienen ciclos dirigidos de orden mayor a uno en *B. subtilis*.

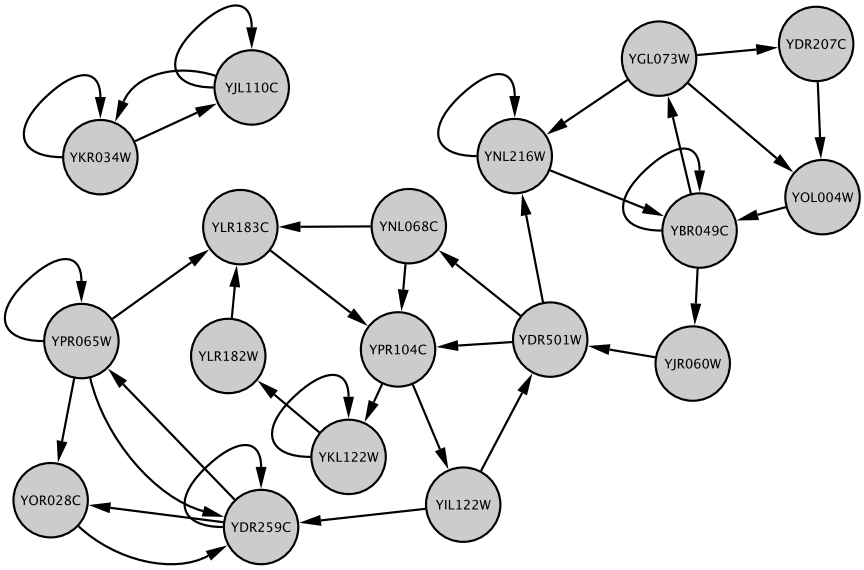

Figura 3.8: Circuitos que contienen ciclos dirigidos de orden mayor a uno en *S. cerevisiae*. Aunque las redes son incompletas, se puede observar un control más intrincado en el eucarionte que en los procariontes. Además, es curioso que la estructura del circuito de dos genes, arriba a la izquierda, se repita en los tres organismos y en la red del desarrollo floral de *A. thaliana* [26].

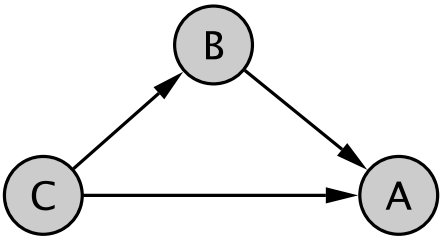

Figura 3.9: Circuito de Alimentación Consecutiva (Feed Forward Loop). Este motivo se encuentra sobrerepresentado en las redes de regulación.

resto (de la red cabeza) jerárquico. Es sólo por los pocos ciclos existentes que no se puede tratar a las redes de regulación como completamente jerárquicas. En total el porcentaje de nodos que participan en ciclos es menor al 7 % en todas las redes de regulación. Esto elimina la duda de una red plagada de ciclos y hace de la inferencia en redes reales un procedimiento correcto.

No se Pueden Determinar Todas las Frases de Regulación

Antes de continuar con el algoritmo de inferencia debemos darnos cuenta que existen frases de regulación que, sin importar el número de condiciones distintas exploradas, son imposibles de inferir. Por ejemplo, supongamos un circuito donde un gen C regula a dos genes A y B y donde, a su vez, el gen B regula al gen A, Fig. 3.9. A este circuito se le conoce como *Circuito de Alimentación Consecutiva* (Feed Forward Loop). Este circuito ha sido recientemente muy estudiado ya que en redes reales de regulación es un motivo que se encuentra sobrerepresentado [63, 64, 65, 66].

Supongamos que C regula negativamente a B. La Tabla de Cambio de B es

C	B
$\downarrow$	$\uparrow$
$\uparrow$	$\downarrow$

Ahora, las posibles combinaciones de aumento/disminución con las que C y B pueden actuar sobre A son

C	B
$\downarrow$	$\downarrow$
$\uparrow$	$\downarrow$
$\downarrow$	$\uparrow$
$\uparrow$	$\uparrow$

A primera vista todas la frases parecen igualmente probables de inferir. Sin embargo, observemos que la primera y cuarta son irrealizables por el tipo de regulación que C ejerce sobre B. Al ser C un inhibidor de B, el que B aumente (disminuya) al aumentar (disminuir) C viola el carácter regulatorio de C sobre B. Seguramente se pueden realizar experimentos que eliminen el control de C sobre B para luego inhibir a ambos o sobreexpresarlos y encontrar cómo estas combinaciones afectan a A. Sin embargo, la red *silvestre*[9] sólo podrá acceder a la segunda y tercera frases anteriormente enlistadas.

Por lo arriba dicho, las frases inferidas serán un subconjunto de las frases que en principio pueden ser determinadas. De esto se desprende que el valor de p se estima sólo con el subconjunto de frases a las que el organismo puede acceder y en este sentido estimamos una p efectiva. Es posible que, si se pudieran conocer todas las frases de regulación, el valor de p de ellas calculado diferiría del valor efectivo. Nosotros creemos que, por las razones biológicas expuestas, el valor efectivo de p es el relevante.

El no considerar en la dinámica frases que son imposibles de inferir facilita el problema de encontrar un primer conjunto de funciones Booleanas para el tipo silvestre de algún organismo ya que la atención se centraría en determinar sólo las frases relevantes. Es claro que las frases que no se pueden determinar en el tipo silvestre son importantes en mutantes o en condiciones experimentales especiales, *e.g.* sobreexpresiones. Sin embargo, una vez conocido el conjunto de frases silvestres, sólo se tienen que determinar las pocas frases que resultan relevantes a los mutantes para luego poder determinar los patrones mutantes de expresión.

En una dinámica, tener en cuenta las frases incompatibles con la lógica de la red, nos permite restringir el conjunto de configuraciones iniciales a sólo aquellas que son compatibles con la coherencia interna de la red. Esto equivale a no considerar como biológicamente relevantes a ramas enteras en las cuencas de atracción[10].

[9] Red silvestre: la que corresponde a la red de regulación de la cepa silvestre.

[10] En una simulación, nada nos prohibe inicializar a la red con una configuración incom-

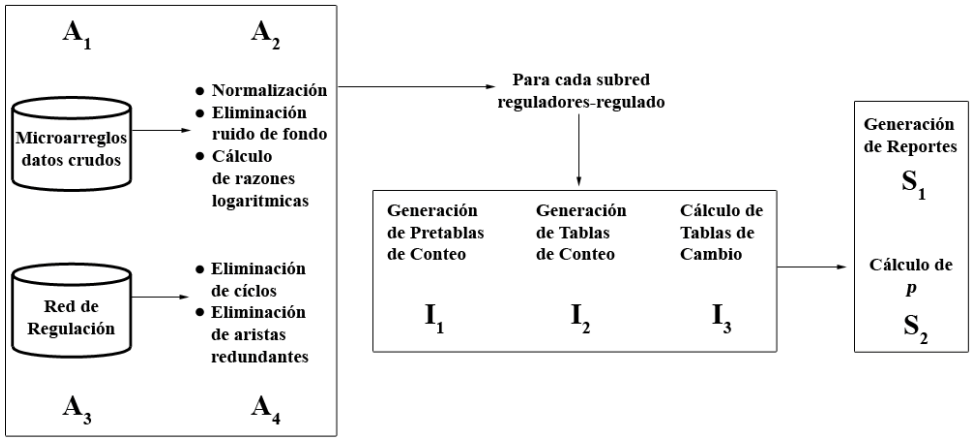

Figura 3.10: Diagrama de flujo que muestra los principales pasos de la inferencia. La primera etapa (A) consiste en adquirir y procesar la red y los microarreglos. La segunda etapa (I) infiere las frases de regulación usando las evidencias en los microarreglos en todas las subredes regulado-reguladores de la red sin ciclos. La última etapa (S) genera reportes donde se indica qué frases fueron inferidas exitosamante, estimando con ellas el valor de p.

3.2.2. El Algoritmo de Inferencia

En la Fig. 3.10 mostramos un diagrama de flujo con los pasos principales de todo el proceso de inferencia. Los pasos se pueden dividir en tres etapas: adquisición y formato de datos (A), inferencia (I) y generación de resultados (S).

A_1 *Alimentamos al algoritmo con datos crudos de microarreglo*[11]. Para cada organismo se eligieron conjuntos de microarreglos de distintas bases de datos. Para *E. coli* todos los experimentos de microarreglo disponibles en

patible con la lógica interna de ésta. El situarnos desde la perspectiva en la que la red trata de encontrar un estado coherente con sus Tablas Booleanas, nos permite reinterpretar la dinámica desde la configuración incompatible como una sucesión de un estado transiente a otro en un proceso en el que la red busca autoconsistencia. Al final, pueden pasar dos cosas, se llega a un atractor puntual completamente consistente con las restricciones impuestas por las Tablas Booleanas o se llega a un ciclo donde ninguna de las configuraciones que lo conforman son autoconsistentes pero que por lo determinista de la dinámica, la red se ve obligada a visitar una y otra vez.

[11] Los datos crudos son aquellos que aún contienen ruido de fondo y no han sido normalizados. Ver Apéndice A

el *Stanford Microarray Database*[12] (SMD) [67] fueron incorporados en el algoritmo excepto aquellos que aumentaron el número de falsos positivos (ver abajo Evaluación del Éxito de la Inferencia); escogimos 107 experimentos en total del SMD. También incorporamos 48 experimentos de [68, 69]. Para *B. subtilis* incorporamos todos los microarreglos disponibles en el *KEGG Expression Database*[13] excepto los experimentos ex0000818 y ex0001438 por aumentar el número de falsos positivos; 69 experimentos en total. Para *S. cerevisiae* incorporamos básicamente todos los experimentos de tres experimentadores en SMD. El nombre de los experimentadores y el número de microarreglos son: Gasch, 138; DeRisi, 29; Spellman, 56. Para una lista completa de identificadores de todos los experimentos incorporados para los distintos organismos ver Apéndice A.

A_2 *Los datos crudos son procesados.* Los datos crudos de microarreglo deben ser procesados para que sean comparables entre si. Los pasos principales consisten en la *eliminación del ruido de fondo*, la *normalización* y el cálculo de las *razones logarítmicas* (base 2). Los detalles de este procesamiento los reportamos en el Apendice A. El procesamiento particular para cada conjunto de microarreglos dependiendo de su fuente es el siguiente. Los datos de SMD los obtuvimos como razones logarítmicas con el ruido de fondo eliminado y normalizados. Todos los datos marcados como no confiables por la base de datos fueron ignorados. Los datos de [68, 69] fueron normalizados y corregido el ruido de fondo usando el programa computacional *Affymetrix Microarray Suite 5.0*. Luego calculamos las razones logarítmicas entre los experimentos en cepas silvestres (con y sin glucosa) y los experimentos de cepas mutantes (con y sin glucosa). Los datos obtenidos del *KEGG Expression Database* ya estaban normalizados al obtenerlos; sólo fue necesario eliminar el ruido de fondo y calcular las razones logarítmicas. Hay que tener en cuenta que, por distintas causas, no se puede obtener de todos los experimentos de microarreglo la intensidade de todos los genes de un organismo. En estos casos, sustituimos la intensidad faltante por -1. El algoritmo de inferencia ignorará a cada evidencia que involucre a un gen/genes con dicho valor en su intensidad.

A_3 *Se alimenta al algoritmo con las redes experimentales.* La red de regulación de *E. coli* se obtuvo de *RegulonDB*[14] versión 5.5 [13]. La red de

[12] `http://genome-www5.stanford.edu`
[13] `www.genome.jp/kegg/expression`
[14] `http://regulondb.ccg.unam.mx`

regulación de *B. subtilis* fue obtenida de *DBTBS*[15] versión 4.1 [14]. La red de *S. cerevisiae* se obtuvo del sitio `http://sandy.topnet.gersteinlab.org` que contiene la información adicional a la publicación [11].

A_4 *Formato de las redes de regulación.* Tuvimos cuidado en eliminar redundancias en las redes de regulación ya que es común que un gen interactue con otro en varios sitios de unión de la región reguladora. Las bases de datos suelen reportar tantas interacciones como sitios de unión. Además, eliminamos de la red todas aquellas regulaciones que fueran a dar a un gen que participara en algún ciclo dirigido. Esto provoca que en el algoritmo de inferencia dichos genes no sean considerados[16]. En la red de *B. subtilis* se consideraron los factores *sigma* ya que, desde el punto de vista dinámico, son elementos equivalentes a los factores de transcripción.

Una vez introducidos los datos repetimos los pasos I_1, I_2 e I_3 para cada subred regulador-regulados que pueda formarse.

I_1 *Generación de pretablas de conteo.* Para cada experimento se obtienen las razones logarítmicas de cambio[17] de los genes en la subred. Las razones logarítmicas se discretizan comparando cada razón con un cambio umbral x_0. Si la razón logarítmica es mayor que x_0 asignamos un aumento en el nivel de expresión. Si la razón logarítmica es menor que $-x_0$ asignamos una disminución en el nivel de expresión. Si la razón cae entre estos dos rangos el cambio es insuficiente para poder ser determinado. Discretizados los cambios, procedemos a contabilizar las evidencias en una tabla que designamos *pretabla de conteo*. Por ejemplo, si la subred fuera $B, C \rightarrow A$ y la evidencia de distintos experimentos, Exp_i, fuera

	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7
A	2.34	1.56	2.05	2.56	0.65	-3.45	-2.55
B	1.76	1.95	2.86	2.67	-1.89	-2.06	-1.79
C	3.36	1.45	1.35	1.97	-1.78	-1.67	-1.99

la evidencia discretizada con x_0 =1.5 sería,

[15] `http://dbtbs.hgc.jp`

[16] Todos los genes no considerados en el proceso de inferencia pertenecen a la cabeza de la red.

[17] La razón logarítmica de cambio de un gen es el logaritmo del cociente entre la intensidad de expresión en el experimento a comparar y la intensidad de expresión en el experimento control. El logaritmo es base dos para saber cuantas veces se ha duplicado/dividido por dos la expresión del gen respecto al control. Ver Apéndice A.

	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7
A	↑	↑	↑	↑	−	↓	↓
B	↑	↑	↑	↑	↓	↓	↓
C	↑	−	−	↑	↓	↓	↓

Con ↓ indicando una disminución en el nivel de expresión, ↑ un aumento y − un cambio insuficiente para poder ser determinado. La evidencia sintetizada en la pretabla de conteo es

		A		
B	C	↓	−	↑
↓	↓	2	1	0
−	↓	0	0	0
↑	↓	0	0	0
↓	−	1	0	0
−	−	0	0	0
↑	−	0	1	2
↓	↑	0	0	0
−	↑	0	0	0
↑	↑	0	0	2

I_2 *Generación de tablas de conteo.* Ignorando todas aquellas instancias de la pretabla de conteo con − y sumando las evidencias *a priori* llegamos a las tablas de conteo. Por ejemplo, al reducir la pretabla anterior y sumar un 1 como evidencia *a priori* a cada opción de cambio de A y en cada combinación de B y C obtenemos,

		A	
B	C	↓	↑
↓	↓	3	1
↑	↓	1	1
↓	↑	1	1
↑	↑	1	3

Notemos que el número de evidencias *a priori* que repartimos en la tabla fue $N = 8$. Para no caer en resultados paradójicos se requiere cumplir con un *tamaño de muestra equivalente* que pide que en cada subred se reparta el mismo número de evidencias entre todas las posibilidades [56, 57]. Esto evita que subredes grandes tengan un *mayor* número de observaciones *a priori* que subredes chicas. El número de evidencias a repartir *a priori* depende del problema y en nuestro caso particular de inferencia de frases de regulación encontramos que son $N = 4$ evidencias *a priori* repartidas

uniformemente entre todas las instancias de la tabla y entre las dos posibilidades del gen regulado[18]. En los siguientes párrafos damos una explicación.

I_3 Cálculo de las Tablas de Cambio. A partir de las tablas de conteo podemos calcular la probabilidad *a posteriori* de que una frase de regulación sea activadora o inhibidora. Esto lo decidimos en base al porcentaje de evidencia en cada frase. Si el porcentaje de evidencia a favor de la activación sobrepasa un umbral t_1 tenemos una activación. Si el porcentaje de evidencia a favor de una activación no sobrepasa un umbral $t_2 < t_1$ tenemos una inhibición. Si el porcentaje se encuentra entre t_2 y t_1 la evidencia no es suficiente para apoyar un tipo u otro de regulación. Supongamos que $t_1 = 70\,\%$ y $t_2 = 30\,\%$ entonces, para la tabla de conteo anterior tenemos,

		A		
B	C	↓	↑	% evidencia activación
↓	↓	3	1	$25\,\% < t_2$
↑	↓	1	1	$t_2 < 50\,\% < t_1$
↓	↑	1	1	$t_2 < 50\,\% < t_1$
↑	↑	1	3	$75\,\% > t_1$

$\Rightarrow$

B	C	A
↓	↓	↓
↑	↓	−
↓	↑	−
↑	↑	↑

siendo la tabla de la izquierda la tabla de conteo con los porcentajes que apoyan una activación y la de la derecha la tabla de cambio inferida; − indica que la evidencia no es suficiente para determinar el tipo de regulación. Nuestra selección del tamaño de muestra equivalente y los umbrales trata de disminuir el número de falsos positivos, es decir, de inferencias que indiquen una activación o una inhibición falsa[19]. Por ejemplo, los parametros $N = 1$, $t_1 = 70\,\%$ y $t_2 = 30\,\%$ provocan muchos falsos positivos ya que basta, *en tablas de genes regulados por un regulador*, una sola evidencia para establecer una frase de regulación.

Para evitar el fenómeno anterior, vamos a determinar que valor de N es necesario para impedir que una sóla evidencia experimental establezca una frase de regulación cuando un gen A es regulado por *sólo* un regulador B,

[18] Por ejemplo, para un gen regulado por tres genes existen 2^3 instancias de la tabla con todos los posibles cambios combinados de los tres reguladores. Para cada instancia de la tabla existen dos posibilidades: el gen regulado aumenta su expresión o la disminuye. De esta forma, la evidencia a priori debe ser repartida en $2 \times 2^3 = 16$ casos; a cada caso le corresponde una evidencia a priori igual a $4/16 = 0.25$ si $N = 4$. Repartir uniformemente la evidencia equivale a suponer una total ignorancia de la regulación.

[19] Ver abajo Evaluación del Éxito de la Inferencia para una explicación de cómo se determinan los falsos positivos.

↓	↑	0.025	0.50	0.80	1.00
0	1	**0.97**	**0.75**	**0.69**	0.66
0	2	**0.99**	**0.83**	**0.77**	**0.75**
1	2	0.66	0.62	0.61	0.60
1	3	**0.75**	**0.70**	0.68	0.66
1	4	**0.80**	**0.75**	**0.72**	**0.71**
2	4	0.66	0.64	0.63	0.62
3	8	**0.73**	**0.71**	**0.70**	**0.69**
4	16	**0.80**	**0.79**	**0.77**	**0.77**
7	23	**0.77**	**0.76**	**0.75**	**0.75**

Cuadro 3.2: Porcentajes que apoyan una activación como función del número de evidencias experimentales (renglones) y del número de evidencias *a priori* (columnas) repartidas en ↑ y ↓, *e.g.* 0.80 corresponde a 0.80 evidencias en ↑ y 0.80 evidencias en ↓. Variando la cantidad de evidencia a priori variamos la importancia relativa de la evidencia experimental. Hacia la derecha de la tabla la importancia de la evidencia experimental disminuye. Se encuentran en negrita aquellos porcentajes que superan el valor umbral $t \geq 0.69$.

$B \rightarrow A$. En el Cuadro 3.2 hemos tabulado porcentajes que apoyan una activación como función de evidencias experimentales y evidencias *a priori*. El número de evidencias experimentales varía a lo largo de una columna y va desde $0 \downarrow 1 \uparrow$ (cero evidencias que apoyan la inhibición y una que apoya la activación) hasta $7 \downarrow 23 \uparrow$. Las evidencias *a priori* varían a lo largo de cada renglón; notemos que no hay nada de malo en asignar evidencias no enteras positivas. Cada elemento de la tabla es el resultado de

$$\text{(evidencia exp.}\uparrow + \text{evidencia } \textit{a priori}\uparrow)/(\text{evidencia exp.}\downarrow \text{ y } \uparrow + \text{evidencia } \textit{a priori}\downarrow \text{ y } \uparrow)$$

Por ejemplo el porcentaje en el renglón tres columna tres es

$$(2 + 0{,}80)/(2 + 1 + 0{,}80 + 0{,}80) = 0{,}61 \tag{3.9}$$

Hemos resaltado en negrita todos aquellos porcentajes que superan el valor umbral $t \geq 0.69$. Observemos que al aumentar la evidencia *a priori* una misma cantidad de evidencia experimental puede llegar a no ser suficiente para apoyar una activación dado el umbral anterior. En particular observemos que una evidencia experimental $0 \downarrow 1 \uparrow$ no es suficiente para concluir una activación con la evidencia a priori de la última columna. La evidencia

a priori de la última columna equivale a un tamaño de muestra equivalente $N = 4$ en la red de un gen A regulado por sólo un regulador B, $B \rightarrow A$, cuya tabla tiene *cuatro* posibles frases de regulación. En el programa de inferencia usaremos este valor de N y además $t_1 \geq 0{,}69$ y $t_2 \leq 0{,}31$. Obviamente podemos ser más estrictos y aumentar N o aumentar/disminuir t_1/t_2, sin embargo esto no es necesario ni deseable ya que el éxito de la inferencia se encuentra alrededor del 90 %[20] y un aumento en los parámetros provocaría que ignorásemos regulaciones correctas.

S_1, S_2 Reportes y cálculo de p. Terminado el proceso de inferencia buscamos todas aquellas tablas de cambio donde haya sido posible infereir al menos una frase de regulación. Para cada una de estas tablas generamos un reporte que detalla cada uno de los pasos antes descritos. Finalmente, con las frases de regulación inferidas estimamos el valor de p (número de activaciones inferidas entre el número total de inferencias).

3.2.3. Resultados de la Inferencia

Antes de estimar el valor de p debemos de alguna forma validar los resultados que obtenemos de la inferencia. Para esto, usamos la información reportada en las bases de datos *RegulonDB* y *DBTBS* sobre el tipo de regulación cuando un gen es regulado sólo por otro. Con esta información, obtenemos las tablas de cambio "reales", mismas que comparamos con las tablas inferidas de redes regulador-regulado. El éxito de la inferencia (coincidencias entre lo inferido y lo reportado) depende del valor umbral x_0. A medida que éste crece, el éxito aumenta mientras que el número de frases inferidas disminuye; determinamos el valor óptimo de x_0 en el que el exito ronda el 90 % y se obtienen cientos de frases. Finalmente, estimamos el valor de p con las frases inferidas en x_0 óptimo y usamos el porcentaje de discrepancias (no-coindidencias entre lo inferido y lo reportado) como una estimación del error en la medida de p.

Evaluación del Éxito de la Inferencia

En las bases de datos de *E. coli* y *B. subtilis* se puede encontrar el tipo de regulación de muchas interacciónes pudiendo ser éstas inhibidoras o activadoras. Esta información es incompleta para formar las frases de regulación. Por ejemplo, para el gen `fecR` que es regulado por `FecI` y `Fur`, se reportan las siguientes interacciones y su tipo:

[20] *Ibid.*

organismo	x_0	**frases comparadas**	**coincidencias**	**activaciones (discrepancias)**	**inhibiciones (discrepancias)**
E. coli	1.50	92	86 %	11 %	3 %
B. subtilis	1.30	36	89 %	11 %	0 %

Cuadro 3.3: Porcentaje de coincidencias entre las frases simples reportadas y las inferidas. La tercer columna indica cuantas frases inferidas fueron comparadas con las reportadas. Las últimas dos columnas son un desgloce del porcentaje de inferencias que discrepan con lo reportado en activaciones (en realidad inhibiciones) e inhibiciones (en realidad activaciones).

$$\begin{array}{lcll} \texttt{FecI} & \rightarrow & \texttt{fecR} & + \\ \texttt{Fur} & \rightarrow & \texttt{fecR} & - \end{array}$$

Al preguntarnos qué sucede cuando `FecI` (activador) y `Fur` (inhibidor) aumentan ambos su nivel de expresión no sabemos si `FecR` aumentará o disminuirá; todo depende de quién sea el represor/activador más fuerte. Uno encuentra el mismo tipo de problema al considerar genes regulados por tres o más reguladores. Sin embargo, cuando un gen sólo tiene un regulador no hay ambigüedad. Como ejemplo, veamos qué sucede cuando el tipo de regulación es positiva (activación):

$$\texttt{Fis} \rightarrow \texttt{alaW} \; +$$

De esta información es inmediato que cuando `Fis` aumenta, `alaW` aumenta y cuando `Fis` disminuye, `alaW` lo sigue:

Fis	alaW
↓	↓
↑	↑

La situación es análoga cuando un gen es regulado por un inhibidor. Por lo tanto, las frases *simples* de regulación se leen directamente del tipo de regulación reportada en las bases de datos.

Para evaluar el éxito de la inferencia hemos comparado todas la frases simples inferidas con las frases simples reportadas. En el Cuadro 3.3 mostramos los resultados obtenidos para *E. coli* con $x_0 = 1{,}5$ y *B. subtilis* con $x_0 = 1{,}30$; estos valores de x_0 los justificamos más adelante. En la tercer columna reportamos el número de frases simples inferidas que comparamos con las reportadas. Este número en *B. subtilis* es inferior al de *E. coli* simplemente porque hay menos frases simples reportadas en *B. subtilis*. En la

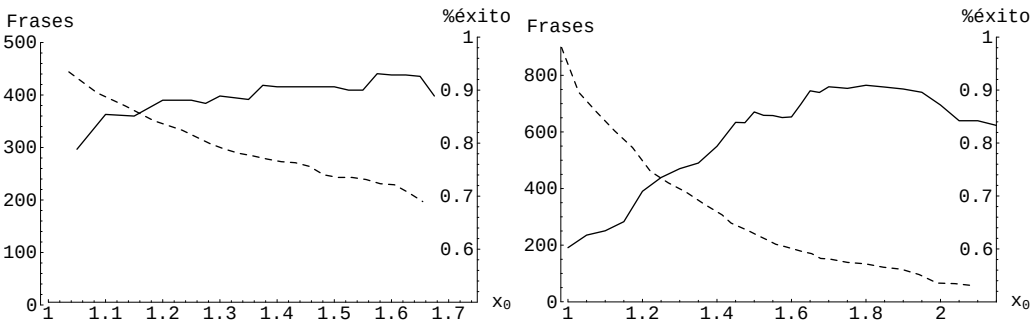

Figura 3.11: Frases inferidas (línea quebrada) y porcentaje de coincidencias (línea continua) ambos como función de x_0. Izquierda *B. subtilis*. Derecha *E. coli*.

cuarta columna mostramos el porcentaje de coincidencias de lo inferido con lo reportado. En la cuarta y quinta columnas separamos el porcentaje de discrepancias en el porcentaje de discrepancias en activaciones y el de discrepancias en inhibiciones.

Notemos que, en el párrafo anterior, no hemos usado la frase *inhibiciones erróneas* para designar a las inhibiciones que discrepan con lo reportado en las bases de datos (análogamente para la frase *activaciones erróneas*). El hecho de que no coincida lo inferido con lo reportado no indica necesariamente que la inferencia sea incorrecta. Una causa posible es el hecho de que las redes experimentales aún son incompletas por lo que es posible que las discrepancias se deban a la existencia de regulaciones aún no encontradas. Por ejemplo, puede suceder que un regulador sea dual y tenga dos sitios de unión a la región reguladora siendo uno de ellos represor y el otro activador. La decisión del lugar al que se va a pegar depende de la condición a la que esté expuesto el organismo. Por lo tanto, la inferencia será incorrecta si sólo se ha reportado una de las interacciones y las condiciones analizadas en los microarreglos favorecen la interacción con el sitio de unión no reportado. De cualquier forma, no se descarta el hecho de que las discrepancias sean en realidad errores en la inferencia debido a, por ejemplo, baja estadística. Teniendo en cuenta las sutilezas anteriores usaremos los porcentajes de las discrepancias para estimar, más adelante, la incertidumbre en el valor de p.

x_0 óptimo

Es de esperar que al aumentar el valor de x_0 el porcentaje de coincidencias aumente y el número total de frases inferidas disminuya. En la Fig. 3.11 podemos notar dicho comportamiento. Se observa una disminución monótona en el número de frases inferidas al aumentar x_0. Por el contrario, el porcentaje de coincidencias inferidas correctamente aumenta aunque después decrece un poco. Este decremento se debe a que los cambios en intensidad de cierto porcentaje de las discrepancias es muy alto y aunque el valor de x_0 aumente, éstas permanecen. En contraste, el aumento de x_0 provoca que frases que coinciden con lo reportado ahora sean ignoradas. Combinados los dos efectos anteriores dan como resultado neto la disminución del porcentaje de coincidencias a partir de cierto x_0.

Deseamos escoger una x_0 tal que tengamos un número relativamente alto de frases inferidas y un buen porcentaje de coincidencias. A continuación presentamos un extracto de la tabla con la que fueron generadas las gráficas de la Fig. 3.11. Hemos resaltado en negrita las líneas con el valor de x_0 que hemos escogido. En *B. subtilis* escogimos x_0 =1.300 ya que el éxito ronda el 89 % y las frases inferidas 307. Antes y después de este valor el éxito disminuye hasta que en x_0 =1.375 el éxito vuelve a aumentar a 91 % pero con 284 frases inferidas. Entre x_0 =1.300 y x_0 =1.375 el éxito aumenta 2 % y el número de frases disminuye 7.5 %. Al escoger x_0 =1.300 hemos optado por una mayor estadistica sacrificando un poco de precision. Por una consideración similar, escogemos x_0 =1.500 para *E. coli*.

	B. subtilis	
x_0	Frases	éxito
1.250	333	0.8710
1.275	320	0.8750
1.300	**307**	**0.8889**
1.325	297	0.8857
1.350	289	0.8824
1.375	284	0.9091
1.400	278	0.9063
1.425	273	0.9063
1.450	271	0.9063
1.475	264	0.9063
1.500	248	0.9063

	E. coli	
x_0	Frases	éxito
1.450	307	0.8390
1.475	278	0.8384
1.500	**264**	**0.8587**
1.525	250	0.8523
1.550	234	0.8519
1.575	219	0.8481
1.600	203	0.8493
1.625	195	0.8732
1.650	186	0.8986
1.675	177	0.8955
1.700	170	0.9062

La determinación del éxito en la inferencia fue posible gracias a que pudimos establecer una relación entre las frases de regulación simples y los tipos de regulación reportados en las bases de datos de *E. coli* y *B. sub-*

tilis. Sin embargo, para *S. cerevisiae* no pudimos encontrar reportados los tipos de regulación y en consecuencia no pudimos establecer el porcentaje de coincidencias. Para poder proseguir, decidimos tomar x_0 =1.500 que es el mayor de los x_0 establecidos por el análisis anterior. Notemos que un umbral x_0 =1.500 equivale a pedir cambios mínimos de 2.83 tantos en las razones logarítmicas (base 2) medidas en los experimentos de microarreglo[21]. Con esto esperamos que para *S. cerevisiae* la inferencia sea tan buena como para los otros dos organismos.

Estimación de p

La estimación del valor de p en base a las frases de regulación inferidas es la culminación de los cálculos hechos hasta el momento. Recordemos que p es la probabilidad de expresión genética y, dada la conexión hecha entre las Tablas Booleanas de Cambio y las de Estado, ésta se estima como el número total de frases activadoras inferidas dividido por el número total de frases inferidas.

Para calcular la incertidumbre[22] en el valor de p usaremos los porcentajes de discrepancias obtenidas en la Cuadro 3.3. En dicho cuadro vemos que las discrepancias se encuentran divididas en activaciones e inhibiciones. Por ejemplo, para *E. coli* se encontró que, del total de discrepancias, un 11 % corresponde en realidad a inhibiciones que fueron inferidas como activaciones; correspondientemente un 3 % son activaciones inferidas como inhibiciones. Con estos valores se puede calcular un límite superior y otro inferior para p de la siguiente forma. Supongamos que en realidad el 11 % de discrepancias en las activaciones resultan ciertas y que el 3 % en las inhibiciones falsas. Para recalcular p tenemos que restar el 11 % de las activaciones y ésta misma cantidad sumarla a las inhibiciones obteniendo un límite inferior de p. Análogamente, el límite superior se obtiene al suponer que el 3 % de las discrepancias en las inhibiciones resultan ciertas y que el 11 % en las activaciones falsas.

Notemos que para *B. subtilis* el porcentaje de discrepancias en las inhibicio-

[21] La razón logaritmica (base 2) de la expresión de un gen es el logaritmo (base 2) del cociente entre el nivel de expresión que se quiere comparar y el nivel control. Un cambio de 2.83 veces en la expresión es aproximadamente $log_2(2^{1,5}) \sim log_2(2{,}83)$.

[22] A la estimación que hacemos de la incertidumbre del valor de p escapa la variación debida a las frases que no se conocen y que podrían cambiar el valor de p más allá de los límites de la incertidumbre aquí calculada.

organismo	x_0	**inferencias totales**	**activaciones**	p	**límite superior**	**límite inferior**
E. coli	1.50	264	152	0.5758	0.5885	0.5124
B. subtilis	1.30	307	163	0.5309	0.5438	0.4723
S. cerevisiae	1.50	196	97	0.4949	0.5505	0.4405

Cuadro 3.4: Valor de p para tres organismos unicelulares. Las últimas dos columnas dan un límite superior y uno inferior al valor de p usando los porcentajes de discrepancias en el Cuadro 3.3.

nes es cero (Cuadro 3.3), por lo que supondremos, para calcular un límite superior de p, que no es cero sino igual al de *E. coli*, 3 %. Para *S. cerevisiae* supondremos que los porcentajes de discrepancia son también idénticos a los de *E. coli*. Los valores estimados para p y las incertidumbres correspondientes se encuentran resumidos en el Cuadro 3.4.

Otra fuente de error en el valor de p reside en el tamaño finito de la muestra con la que hacemos la estimación. Para poder cuantificar este error, generamos varias muestras con tantas frases de regulación como hayamos inferido usando un valor de probabilidad de expresión genética $p_0 \neq p$. Hecho esto, calculamos qué fracción de todas las muestras presentan una fracción de activaciones idéntica a la p que hemos determinado con cierta tolerancia. Por ejemplo, para *E. coli* calculamos para valores de p_0 que varían desde p_0 =0.4258 hasta p_0 =0.7258 con incrementos de 0.005 qué fracción de 100,000 muestras de 264 frases presentan una fracción de activaciones dentro del rango $(p - 0{,}005, p + 0{,}005)$ con p =0.5758, ver Cuadro 3.4. En la Fig. 3.12 hemos graficado lo anterior como la probabilidad $P(p_0)$ de que una muestra generada con una p_0 distinta de la p calculada presente una fracción de activaciones idéntica a p dentro de una tolerancia de $\pm 0{,}005$ para la p de *E. coli*. La distribución es Gaussiana (línea continua en la gráfica) con un promedio p =0.5758 y una desviación estandar de σ =0.0304. Las gráficas son similares para *B. subtilis* y *S. cerevisiae*. En el Cuadro 3.5 sintetizamos nuevamente los valores de p obtenidos pero con el error debido al tamaño finito de la muestra de tal forma que los límites abarcan el 80 % de los valores de p_0 que con mayor probabilidad generan una fracción de activadores similar a p en una muestra arbitraria, *i. e.* ± 1.24 desviaciones estandar. Notemos que estos rangos de tolerancia se traslapan con los calculados en el Cuadro 3.4.

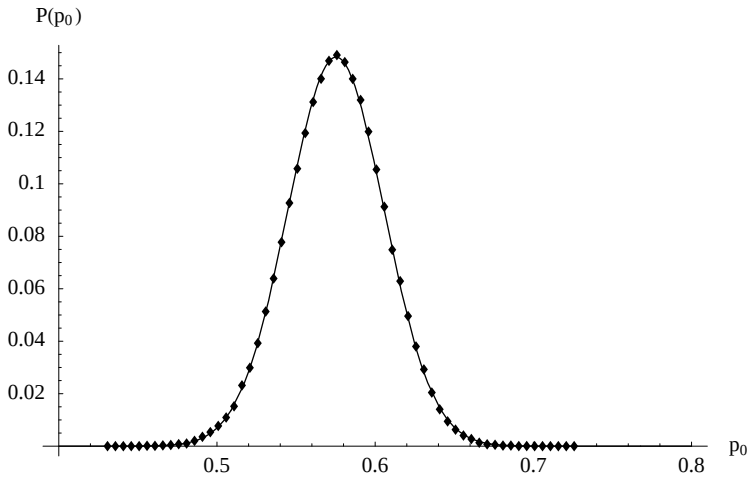

Figura 3.12: Probabilidad $P(p_0)$ de que una muestra de 264 frases generada con una $p_0 \neq p$ contenga una fracción de regulaciones positivas identica a p con una tolerancia de $\pm 0{,}005$, con p la probabilidad de expresión genética estimada para *E. coli*. Notemos que esta probabilidad es Gaussiana (línea continua).

organismo	p	$\pm\mathbf{1.4}\sigma$
E. coli	0.5758	0.0377
B. subtilis	0.5309	0.0354
S. cerevisiae	0.4949	0.0440

Cuadro 3.5: Valor de p para tres organismos unicelulares. El error en la estimación de p es debido exclusivamente al tamaño finito de la muestra (± 1.4 desviaciones estandar $\sim 0{,}80\,\%$ del area de la Gaussiana). Notemos que los rangos de error son similares a los mostrados en el Cuadro 3.4.

organismo	**genes en la red**	**interacciones**	K	p
E. coli	1,328	2,822	2.125	0.576
B. subtilis	830	1,267	1.527	0.531
S. cerevisiae	3,459	7,074	2.045	0.495
A. thaliana	15	43	2.867	0.532
D. melanogaster	60	144	2.4	0.375

Cuadro 3.6: Síntesis de valores que caracterizan a cinco redes regulatorias.

3.3. Organismos Críticos

Hemos obtenido el valor más probable de p para tres organismos unicelulares. En la literatura, además, podemos encontrar otras dos redes: la red responsable del desarrollo floral en *Arabidopsis thaliana* [26] y la red responsable de la generación de los segmentos polares en *Drosophila melanogaster* [25]. Estas redes son muy chicas comparadas con las estudiadas hasta ahora, sin embargo, tienen la ventaja de que todas las funciones Booleanas de los genes participantes han sido determinadas en base a hechos experimentales. El Cuadro 3.6 sintetiza la información más relevante de las 5 redes.

Recordemos que en la primera sección mostramos cómo, con la curva de Derrida, es decir la gráfica del mapeo M, es posible caracterizar la fase dinámica en la que opera una red. En particular, mostramos cómo obtener M de una red infinita con genes independientes y estadísticamente equivalentes, Ec.3.5. Para obtener el mapeo M de las redes reales, debemos considerar dos casos: redes de las que conocemos todas las funciones Booleanas (red de *A. thaliana* y *D. melanogaster*) y redes en las que sólo conocemos p (*B. subtilis*, *E. coli* y *S. cerevisiae*). Para las primeras, la definición de M es inmediata. Para las segundas hemos generado conjuntos de funciones aleatorias acordes con la p calculada. A causa de nuestra total ignorancia de las funciones Booleanas, ésta es la hipótesis nula más adecuada. Para calcular la curva de Derrida de las redes reales, generamos una configuración al azar y perturbamos a una segunda configuración respecto a la primera de tal forma que estén alejadas una distancia de Hamming $D(0)$. Después, evolucionamos bajo la dinámica de la red a ambas configuraciones un paso y volvemos a medir la distancia de Hamming para obtener $M = D(1)$.

En la Fig. 3.13 hemos graficado las curvas de Derrida calculadas numéricamente para los 5 organismos. La recta gris en línea punteada corresponde a

la identidad. La curva de Derrida de las redes reales se muestra con barras de error de una desviación estandard. Para las redes de los microorganismos las barras de error ya consideran el error en la estimación de p además de considerar el error debido a la distribución de las 20,000 perturbaciones que se hicieron para calcular cada punto. Para la planta y la mosca, las barras de error se deben únicamente a la distribución de las 20,000 perturbaciones. Notemos que M llega casi tangente a la identidad en $D = 0$. Esto indica que los datos actuales son compatibles con que las redes de regulación operan alrededor de la fase crítica (comparar con la gráfica 3.2).

Para conocer el valor de la sensibilidad, *i.e.* la pendiente con que llega M al origen, hemos ajustado un polinomio. Los cálculos de la primera sección muestran que M, Ec. 3.5, es un polinomio de grado igual al máximo de reguladores por gen. Hemos usado este hecho como criterio para decidir el grado del polinomio ajustado. La pendiente (sensibilidad) en $D = 0$ del polinomio ajustado para los distintos organismos es: S =1.081 para *E. coli*, S =1.036 para *S. cerevisiae*, S =0.826 para *B. subtilis*, S =0.914 para *D. melanogaster* y S =1.127 para *A. thaliana*. En todos los casos, el coeficiente de regresión fue menor a 10^{-4}, lo que indica un excelente ajuste. Hemos graficado estos valores para tener una mejor idea de cómo se distribuyen alrededor de $S = 1$, Fig. 3.14. Podemos observar que, en base a la división dinámica de referencia obtenida usando el *ensemble* de redes aleatorias, *las redes reales se distribuyen alrededor de* $S = 1$ *lo que indica una dinámica crítica.*

3.4. Conclusión.

Criticalidad, Característica Genérica de la Regulación

En los párrafos anteriores calculamos la curva de Derrida de la red de regulación de cinco organismos que cubren 4 reinos de la vida: el animal *Drosophila melanogaster*, la planta *Arabidopsis thaliana*, el hongo *Saccharomyces cerevisiae* y las bacterias *Escherichia coli* y *Bacillus subtilis*. En todos ellos encontramos que la dinámica de la red regulatoria opera alrededor de la fase crítica. Encontrar en organísmos tan disímiles el mismo comportamiento dinámico sugiere que la criticalidad podría ser una característica genérica de los organismos a nivel genético.

Existen otros indicios que sustentan la idea de que las redes de regulación

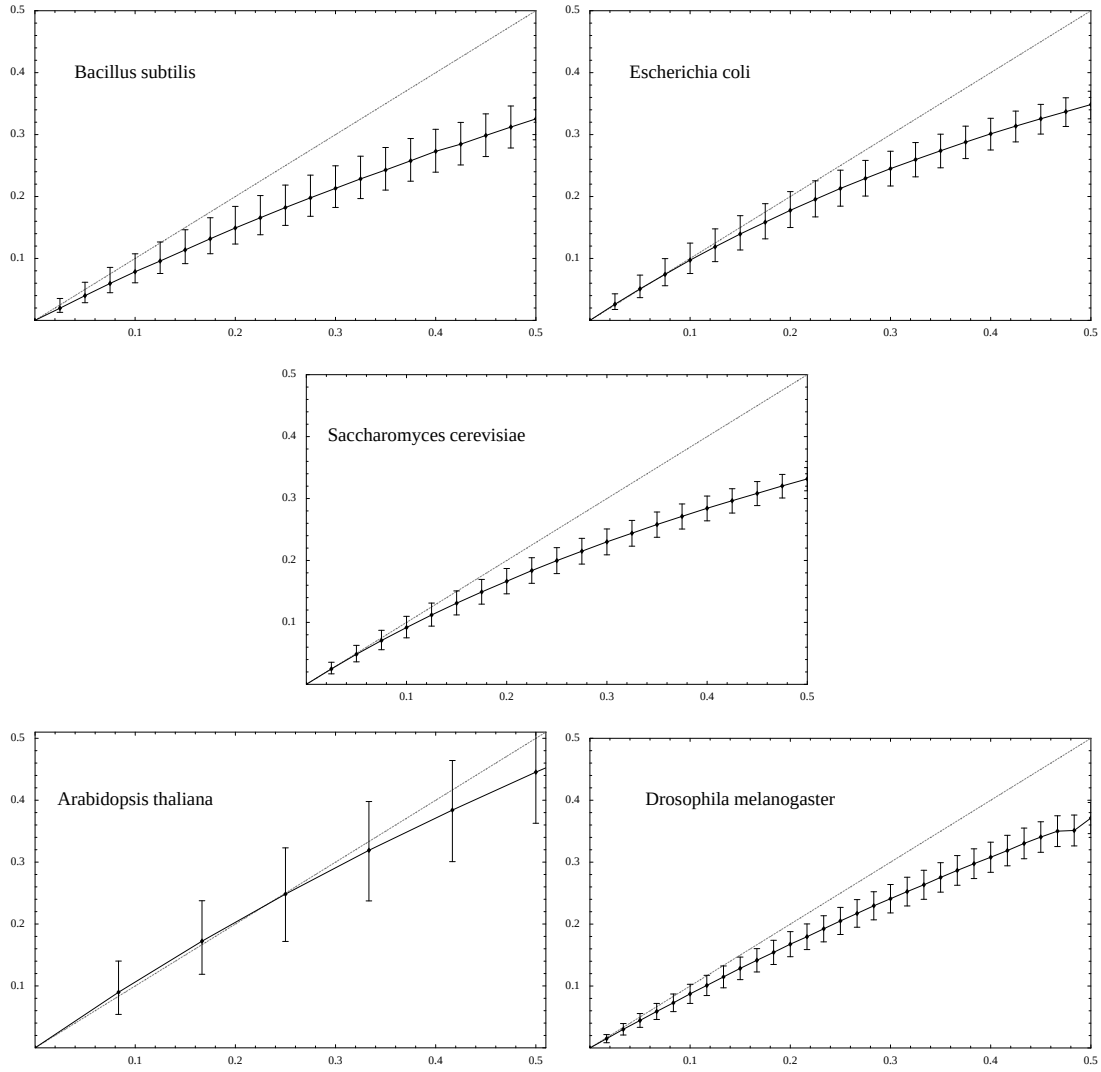

Figura 3.13: Curvas de Derrida de redes de regulación de cinco organismos. La recta corresponde a la identidad. M es la curva de Derrida obtenida directamente de las redes reales. Cada punto es el promedio de 20,000 perturbaciones. Las barras de error son una desviación estandar del promedio del valor obtenido de las 20,000 perturbaciones. Además, para los tres microorganismos, las barras de error también consideran el error debido a la estimación de p. Notemos que M llega casi tangente a la identidad en $D = 0$ lo que indica que las redes reales operan alrededor de la fase crítica.

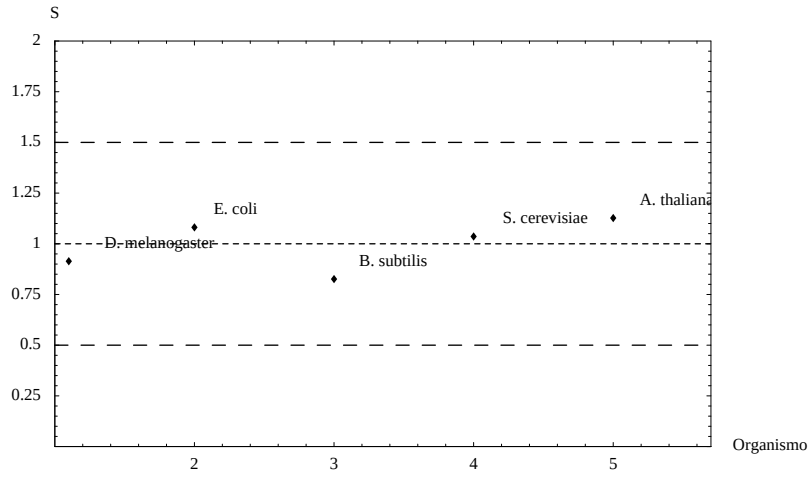

Figura 3.14: Sensibilidad de 5 organismos. Para redes críticas es igual a uno. La sensibilidad de los organismos aquí analizados se encuentra al rededor de ese valor. Para comparar, la línea discontinua inferior en 0.5 es la sensibilidad de redes ordenadas con $p = 1/2$ y $K = 1$. La línea discontinua superior es la sensibilidad de redes caóticas con $p = 1/2$ y $K = 3$

operan alrededor de la fase crítica. Sin embargo, éstos son indirectos o se basan en características cualitativas del modelo Booleano que no han sido corroboradas experimentalmente. Por ejemplo, redes ordenadas exhiben un pequeño número de atractores con transientes pequeños. Por el contrario, redes caóticas codifican un gran número de atractores y el número de configuraciones transientes es enorme [53, 55]. Las redes críticas están en medio de estos comportamientos extremos y uno esperaría que las redes reales así se comportaran. Por otra parte, existe el hecho experimental de que los tamaños de avalanchas provocados por deleciones de genes en *S. cerevisiae* siguen una distribución Libre de Escalas. Sólo redes críticas –que no consideran la topología particular de la red de levadura– pueden reproducir este resultado experimental [70, 71]. Como otro indicio indirecto de la cercanía a la criticalidad, se ha calculado la variabilidad de patrones de expresión de células cancerosas. Ésta es sólo compatible con lo que se obtiene de redes ordenadas o críticas pero no de lo que se obtiene de caóticas [72]. Resaltamos que el presente trabajo *es el primero en medir directamente el estado dinámico de un conjunto de varios organismos sin ignorar su topología, encontrando que éstos operan alrededor de la fase crítica.*

Es claro que las redes reales –que son finitas– nunca van a caer exactamente en la transición de fase, pero mientras más cerca estén, más probable será que transmitan señales de forma coherente sin perturbar a toda la red. Recordemos que la curva M se calcula promediando la distancia entre miles de pares de configuraciones ligeramente distintas una de otra después de evolucionarlas un paso. En redes *finitas ordenadas* la gran mayoría de las configuraciones convergen aunque algunas pocas pueden llegar a divergir y otras tantas conservar su distancia inicial. Esto se traduce en una transmisión de señales y en una respuesta de cambio pobre. Análogamente, en redes *caóticas finitas* la gran mayoría de las configuraciones divergen habiendo unas cuantas que convergen o conservan su distancia: estas redes son poco robustas y la probabilidad de transmitir una señal sin que afecte a toda la red es baja. Sin embargo, al ser la transición de fase continua, mientras más cerca estén las redes de la fase crítica, una mayor proporción de pares de configuraciones ligeramente distintas conservaran su distancia a lo largo de la dinámica. Esta proporción es máxima en la transición de fase.

Decir que una red es crítica implica que tiene la capacidad de cambiar siguiendo las señales ambientales sin ser demasiado sensible como para que el ruido pueda afectar la dinámica de la red. Debemos aclarar que no mostramos específicamente cuándo o bajo qué condiciones la red cambia o permanece estable. Lo que mostramos es que las redes ordenadas permanecen fijas ignorando su ambiente y las caóticas cambian bajo la acción de cualquier señal espuria. Sólo las redes críticas tienen la clara ventaja evolutiva de ser robustas y adaptables.

Siempre es polémico discurrir *en general* al rededor del asunto de si existe progreso en la evolución. Dobzhansky [73] define el progreso como

> el cambio sistemático de una característica que se presenta en todos los miembros de una secuencia de tal modo que los miembros posteriores muestren una mejora de dicha característica.

Los miembros de la secuencia son los organismos y la secuencia la constituye la historia fósil. La mejora es una mejora estadísticamente significativa (que acepta fluctuaciones) a lo largo de la secuencia. Dobzhansky cita como posible criterio de progreso *general* en todas la líneas evolutivas a lo largo de la vida a la expansión de la vida misma (tendencia a ocupar todos los espacios disponibles de los ambientes habitables, incluidos los creados por este mismo proceso de expansión) que se puede medir por el número de tipos

de organismos, el número de individuos y la cantidad de materia viva. También menciona el progreso *particular* en líneas específicas durante algunos periodos de tiempo: mejora en la adaptación, en la adaptabilidad, mayor especialización, mayor control sobre el ambiente, mayor complejidad estructural, incremento en el margen y variedad de ajustes, entre otros. Siguiendo a Dobzhansky, y aceptando la hipótesis de que la robustez dinámica y la adaptabilidad de las redes de regulación son características ubicuas y por tanto escenciales de todo ser vivo, podemos especular sobre una tendencia general de los organismos hacia la criticalidad. Esto genera inmediatamnete las siguientes cuestiones: ¿las redes de regulación primitivas eran críticas cuando surgieron o eran caóticas u ordenadas con una tendencia hacia la criticalidad a lo largo de la evolución de la vida? y ¿existen mecanismos biológicos generales que causen el acercamiento a la fase crítica de las redes de regulación? Las respuestas a estas preguntas abren nuevas líneas de investigación.

Capítulo 4

Robustez y Evolucionabilidad

La duplicación y divergencia genética es un proceso que provoca el crecimiento del genoma y la aparición de nuevas funciones[1] [74, 75]. En casi la totalidad de la literatura existe un sesgo en el que *novedad funcional* se traduce en *una proteína con una nueva función* [76, 77, 78, 79]. No sólo eso, muchas veces se da por sentado que la integración de las nuevas proteínas (funciones) sucede al final de cierto proceso celular, de forma tal que la aparición de la nueva función no *interfiere* con las funciones ya establecidas. Por ejemplo, Lynn Margulis, al explicar cómo han surgido y se han diversificado las vías metabólicas [80], señala que las fases finales de las vías son las de aparición más reciente en la historia evolutiva –evitando con esto la interferencia en la parte de la vía anteriormente establecida.

La utilidad de pensar a una nueva función como una nueva proteína y la tendencia a pensar las vías metabólicas como cadenas causa-consecuencia es indudable. Sin embargo, esto no ayuda a esclarecer hechos como la pleiotropia, la promiscuidad en vías metabólicas[2] [81] o el que un genoma cuyos genes posean regiones amplias de regulación puedan llevar a cabo programas más intrincados de desarrollo[3]. Un ejemplo particularmente sorprendente es

[1] El genoma puede crecer no sólo por duplicación, sino también por transferencia horizontal. Esta consiste en la transferencia de material genético de un organismo a otro excluyendo la transferencia por mecanismos hereditarios (transferencia vertical).

[2] Es sólo recientemente que se ha reconocido que las vías metabólicas estan llenas de entrecruces y no son sólo una colección de vías aisladas. En las redes metabólicas, existen distintas enzimas *promiscuas* que causan la interconexión.

[3] La adquisición de *nuevos programas* codificados por un conjunto de genes puede ser

el hecho de que muchas proteínas pueden ser asociadas con procesos celulares contradictorios como lo es la división celular y la apoptosis [32]. Citando a Huang [81]:

> A pesar de los grandes esfuerzos invertidos en la anotación funcional, sólo en casos excepcionales una función fisiológica específica puede ser *incondicional* e *inambiguamente* asignada a una proteína si ésta es considerada como entidad aislada, definida sólo por la región codificante de su gen correspondiente.

En este capítulo cambiamos el énfasis que se da a la duplicación/divergencia como fuente de proteínas con funciones novedosas haciendo ahora hincapié en que el mismo proceso puede dar lugar a una neofuncionalización que involucra a todos los genes en la red dando origen a nuevos estados funcionales. Esto es posible gracias a que la red genética es *robusta y evolucionable bajo la adición de un nuevo gen*. Mostramos ésto al caracterizar una red (encontrar sus atractores) y posteriormente añadir un nuevo gen, lo que provoca que el paisaje de atractores cambie. Esto es una caricatura del proceso de duplicación genética seguida de divergencia del duplicado. El experimento anterior puede tener varios efectos sobre los atractores de la red original (antes de que suceda la duplicación-divergencia) pudiendo éstos cambiar, desaparecer o permanecer idénticos. Algo inesperado es que también *aparecen nuevos atractores sin que ésto signifique modificar el conjunto original*. Es esta capacidad de integración de nuevos patrones al acervo anterior lo que identificamos como robustez y evolucionabilidad. Mostramos que esta propiedad es óptima en redes críticas independientemente de la topología de éstas.

---

más importante que la simple adquisición de *nuevos genes* para lograr una mayor organización en los sistemas vivos. Un ejemplo de lo anterior se encuentra al comparar los genomas de *Drosophila melanogaster* y *Caenorhabditis elegans*. La mosca tiene menos genes que el gusano –alrededor de 14,000 contra 19,000. Por otra parte, respecto al gusano, la mosca tiene casi el doble de ADN no codificante por gen –alrededor de 10,000 nucleótidos en promedio contra 5,000. A pesar de que hay menos genes en la mosca que en el gusano, el primero parece ser más complejo que el segundo. Alberts *et al* [82] explican que, en la mosca, *el equipo de construcción molecular tiene menos tipos de partes, pero las instrucciones de ensamblaje, especificadas por las secuencias regulatorias en el ADN no codificante, parecen ser más voluminosas.*

4.1. Consideraciones Teóricas

La robustez y la evolucionabilidad son dos propiedades centrales de los sistemas biológicos [83, 84, 85, 86, 87, 88]. Los organismos vivos son robustos ya que pueden realizar sus funciones bajo un amplio rango de pertubaciones azarosas que van desde cambios químicos o físicos transitorios en el ambiente hasta mutaciones genéticas permanentes. También son evolucionables, ya que eventualmente cambian como resultado de modificaciones en su material genético, adquiriendo nuevas funciones y adaptándose a nuevos ambientes. Existen diversas definiciones de estas dos propiedades que dependen del contexto y del nivel de organización considerado. Aquí, para robustez, adoptamos la definición de de Viser *et al.* [84]: *robustez es la invariancia del fenotipo ante perturbaciones.* Igualmente seguimos a Wagner en su definición de evolucionabilidad [89]: *Un sistema biológico es evolucionable si puede adquirir nuevas funciones (fenotipos) a traves de cambio genético, funciones que ayuden al organismo a sobrevivir y reproducirse.* A continuación nos avocamos a definir en las redes Booleanas los conceptos de robustez y evolucionabilidad bajo el mecanismo de duplicación-divergencia.

Se pueden estudiar distintos aspectos de la estabilidad dinámica de las redes Booleanas. Por ejemplo, cuando el sistema se encuentra en un atractor podemos provocar una perturbación (cambiar el estado de algunos genes en la configuración sin cambiar a las funciones Booleanas) que haga al sistema saltar hacia algún estado transiente. Después de la perturbación, el sistema puede o no regresar al atractor inicial dependiendo del tamaño de su cuenca. En particular en redes críticas y ordenadas el sistema regresa al atractor original con alta probabilidad. Reconocemos la capacidad del sistema para regresar al atractor inicial como un tipo de estabilidad dinámica[4]. Otro aspecto de la estabilidad de la red que se puede estudiar es cómo se reconfigura el paisaje de atractores al modificar una función Booleana. Esto simula, por ejemplo, la modificación por mutación de la regulación que experimenta un gen. En modelos de redes reales es común explorar la estabilidad dinámica de la red bajo esta perturbación encontrando que hay una gran probabilidad de que los atractores se conserven sin cambios [26, 90].

En este capítulo estamos interesados en la estabilidad del paisaje de atractores ante la perturbación adición de un gen. En general, la perturbación

[4]Otra forma de ver lo anterior es que el sistema alcanza el estado estacionario "correcto" aún sin especificar de forma precisa el estado inicial del que parte.

modifica a las cuencas de atracción y a los atractores pudiendo éstos últimos desaparecer o, inesperadamente, surgir. Dada la correspondencia entre atractores y tipos celulares/estados funcionales (ver Cap. 2), ciertas modificaciones al paisaje no serán ventajosas para el organismo. Por ejemplo, la pérdida de un atractor por la adición de un nuevo gen significa que algún tipo celular se he inestabilizado y por ende perdido. Esto nos guía hacia la siguiente definición de *robustez en redes Booleanas*:

> *Una red Booleana es robusta ante la adición de un nuevo gen si el paisaje conserva todos sus atractores.*

Haciendo uso de la definición anterior definimos *evolucionabilidad en redes Booleanas*:

> *Una red es evolucionable si, además de conservar sus atractores, el paisaje puede adquirir nuevos al añadir un nuevo gen.*

Estas definiciones sólo conciernen a los atractores y no a su cuenca de atracción que, como hemos mencionado, podría cambiar.

El entender cómo se modifica el paisaje de atractores al agregar un gen, puede ampliar nuestro entendimiento sobre cómo las células/organismos aumentan su repertorio de estados funcionales/fenotipos a través del mecanismo de duplicación-divergencia. Agregar un nuevo elemento a la red podría parecer no estar directamente relacionado con dicho mecansimo. Para mostrar la relación podemos imaginar el siguiente escenario: Supongamos que hemos caracterizado una red de regulación *inicial*, *i.e.* conocemos sus atractores o los patrones de expresión que codifica. Ahora supongamos que sucede la duplicación de alguno de sus genes y que *dejamos de observar a la red por cierto tiempo*. Durante tal periodo, alguna de las copias del gen duplicado divergirá. Al volver a observar a la red y caracterizarla puede suceder que los nuevos patrones de expresión difieran de los originales. ¿Con qué frecuencia la red es capaz de adquirir nuevos patrones de expresión?, ¿Con qué frecuencia la perturabación deja invariantes los patrones originales?. Notemos que dejamos de observar a la red durante cierto tiempo para enfatizar el hecho de que no estamos interesados en saber *cómo* diverge el gen duplicado –cosa que es de lo más interesante en sí misma y actualmente sujeta a intensos estudios [76, 75, 91].

4.2. El Modelo

En el capítulo anterior vimos que las redes de regulación tienen una *cabeza* que codifica distintos patrones de expresión para coordinar la acción de genes estructurales en tareas específicas. Observemos que, de todos los eventos de duplicación/divergencia, los más importantes (desde el punto de vista del control celular) son los que ocurren al nivel de esta *cabeza*. Intuitivamente, estos cambios parecieran ser los más improbables. Sin embargo, los hechos muestran que en una red genética, la parte que lleva a cabo el control es la más *plástica*, *i.e.* tanto la secuencia como los sitios de unión de los factores de transcripción están poco conservados entre organismos en comparación a la conservación de proteínas estructurales [91, 92]. Esto indica que, entre organismos, la interacción específica proteína-proteína/enzima-sustrato es estereotipada. La variedad e innovación radica en cómo, cuándo y con qué intensidad se usan estas interacciones estereotipadas –mucha de la *creatividad* de los organismos se da a nivel regulatorio.

Por lo anterior, sólo consideramos el efecto de la duplicación-divergencia en los Factores de Transcripción (FT). Un duplicado de un FT puede sufrir las siguientes divergencias: (i)divergencias que provocan un cambio en la estructura del FT que a su vez provoca cambios en cómo regula a sus genes objetivo. (ii)Divergencias en la estructura del FT que provocan cambios en los sitios de unión que reconoce. (iii)Cambios en la región de regulación del FT que modifican cómo éste es regulado y (iv)cambios en la región de regulación del FT que modifican por quién es regulado el FT. Las divergencias del FT en (i) corresponden a cambios en la función Booleana de los genes objetivo. Los cambios en (ii) son un recableado de la red que modifica el conjunto de genes objetivo del FT. En (iii) tenemos cambios en la propia función Booleana del FT y en (iv) tenemos un recableado de los genes cuyo objetivo es el FT, ver Fig. 4.1.

De todos los cambios que un FT duplicado puede sufrir, consideramos la combinación que posiblemente afecta de forma más drástica a la dinámica de la red: la ocurrencia simultanea de los cuatro tipos de divergencia arriba expuestos. Considerar sólo la forma más extrema de divergencia permite imponer un limite inferior a la robustez que posiblemente se observaría en casos menos extremos. Todos los cambios que implica esta divergencia extrema son completamente aleatorios, *i.e.* el conjunto de nuevos reguladores/regulados del FT que ha divergido no depende en forma alguna del conjunto inical de reguladores/regulados; lo mismo aplica para la función Boolena del FT que

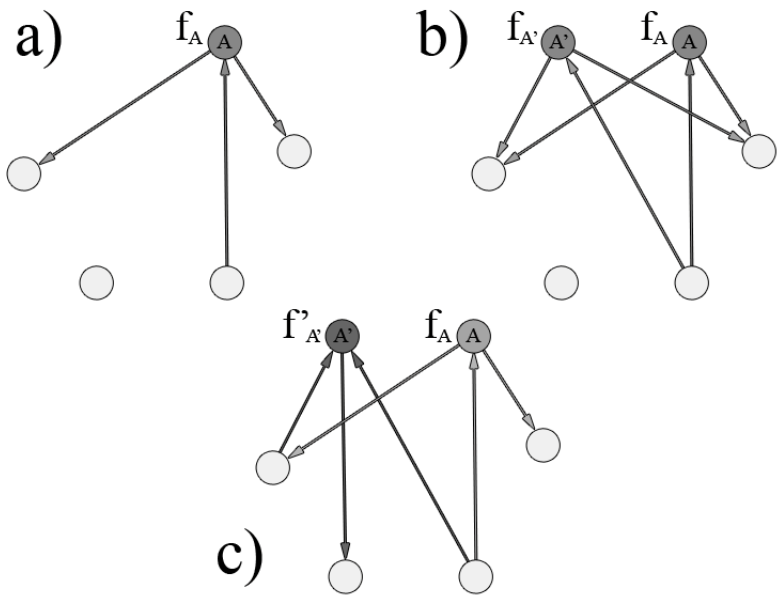

Figura 4.1: Modelo de duplicación genética y divergencia. **a)** Red antes de sufrir duplicación. Sólo dibujamos las regulaciones del gen A, mismo que duplicaremos. **b)** Denotamos al duplicado del gen A por A'. Este gen regula a los mismos genes y tiene los mismos reguladores que el gen A, también la función Booleana $f_{A'}$ es identica a f_A. **c)** Después de la divergencia, el gen A' ha adquirido nuevos reguladores y genes regulados, además de cambiar también su función Booleana que ahora pasa a ser $f'_{A'}$.

ha divergido. Sin embargo, hemos tenido cuidado de extender las funciones Boolenanas de los regulados por el FT que ha divergido de tal forma que cuando el FT no está expresado (estado del FT idéntico a 0) las funciones Boolenas de los regulados son idénticas a las funciones Boolenas previas a la divergencia del FT. En caso de que el FT esté expresado (estado idéntico a 1) el estado de los regulados por el FT que ha divergido se decide al azar, ver Cuadro 4.1.

Cuando añadimos un nuevo gen a la red lo conectamos a ésta siguiendo las características topológicas de la misma. Nuestro estudio gira alrededor de redes con dos topologías: redes con la topología del *Modelo NK* y redes con conectividad de salida Libre de Escala y de entrada Poissoniana. Cuando

B	C	f_A
0	0	0
1	0	1
0	1	0
1	1	1

$\longrightarrow$

B	C	D	f_A
0	0	0	0
1	0	0	1
0	1	0	0
1	1	0	1
0	**0**	**1**	**1**
1	**0**	**1**	**0**
0	**1**	**1**	**0**
1	**1**	**1**	**1**

Cuadro 4.1: **Izquierda:** Antes de la duplicación-divergencia, el gen A sólo es regulado por B y C; su función Booleana consiste de cuatro valores. **Derecha:** Después de la duplicación-divergencia es posible que el nuevo gen D regule al gen A. De ser este el caso extendemos la función Booleana f_A para tomar en cuenta el efecto de D sobre A. Observemos que, cuando D no está presente ($D = 0$), la función extendida es identica a la función original. Cuando D se expresa ($D = 1$) los valores que toma f_A son asignados al azar (parte de la tabla en negritas).

añadimos un gen a redes del *Modelo NK* su conectividad de entrada es K y su conectividad de salida la seleccionamos de una ditribución de Poisson con promedio K. Cuando añadimos un gen a redes con la segunda topología, su conectividad de salida la seleccionamos de una distribución Libre de Escalas con promedio K (ver más adelante) y su conectividad de entrada de una distribución de Poisson también con promedio K. Los K reguladores y los K regulados se eligen al azar de entre todos los genes en la red. Notemos que la forma en la que añadimos un nuevo gen aumenta un poco el promedio de la conectividad de la red que sufre la duplicación-divergencia.

En lo subsecuente llamaremos *red original* a la red previa a la duplicación-divergencia y *red mutada* a la red que ha sufrido dicha perturbación. Reiteramos que la caricatura que proponemos ignora cómo ocurre una duplicación-divergencia específica. Lo que realmente nos interesa saber es cómo reacciona en general la dinámica de la red genética ante dicha perturbación.

4.3. Conservación e Innovación

Todos los resultados que a continuación se presentan se obtuvieron de redes originales con 20 genes y redes mutadas con 20+1 genes promediando sobre 20,000 realizaciones las cantidades relevantes. El valor de la probabilidad de expresión genética se fijó en $p = 1/2$. De esta forma, si K es el promedio de reguladores por gen, las redes ordenadas se situan por debajo de $K = 2$, las redes caóticas por arriba de $K = 2$ y las redes críticas en $K = 2$. Comenzamos discutiendo cómo caracterizar los cambios en los atractores. A continuación exploramos redes con topología homogénea (*Modelo NK*) dónde todos los genes tienen el mismo número de reguladores. Después exploramos redes con topología Libre de Escala en la conectividad de salida. Es importante considerar redes con la primer topología por su importancia en la literatura. Éstas han sido estudiadas desde principios de los setentas y se conocen bien varias de sus propiedades [24, 53, 19, 93]. La importancia de la segunda topología radica en que es la más cercana a la de redes de regulación reales [21, 12], ver Apéndice B.

4.3.1. Los atractores se transforman

Como ya hemos mencionado, el conjunto de atractores representa el conjunto de tipos celulares/estados funcionales de una célula. La perturbación duplicación-divergencia modifica –en general– el conjunto de atractores de la red original aumentando su cardinalidad, disminuyéndola o dejándola idéntica en la red mutada. Biológicamente, sólo es relevante estudiar aquellas redes mutadas que contengan igual o mayor número de atractores que la original ya que redes mutadas con un menor número de atractores *no tienen ni siquiera la posibilidad de conservar al conjunto original*. Por lo tanto, antes de intentar caracterizar los cambios de los atractores, debemos saber con qué probabilidad el número de atractores en la red mutada es mayor o igual que en la original. La gráfica de la Fig. 4.2 muestra precisamente ésto como función del número de reguladores promedio K para redes con topología homogénea (*Modelo NK*) y redes con topología cercana a la observada en redes reales (conectividad de salida Libre de Escala y conectividad de entrada Poissoniana), véase más adelante. Podemos observar que la probabilidad de tener mayor o igual número de atractores disminuye conforme aumenta K y que la probabilidad no es inferior a 0,6 (recordemos que las redes de regulación se encuentran alrededor de $K = 2$). No debe extrañar que esta probabilidad sea relativamente alta para cualquier K *independientemente*

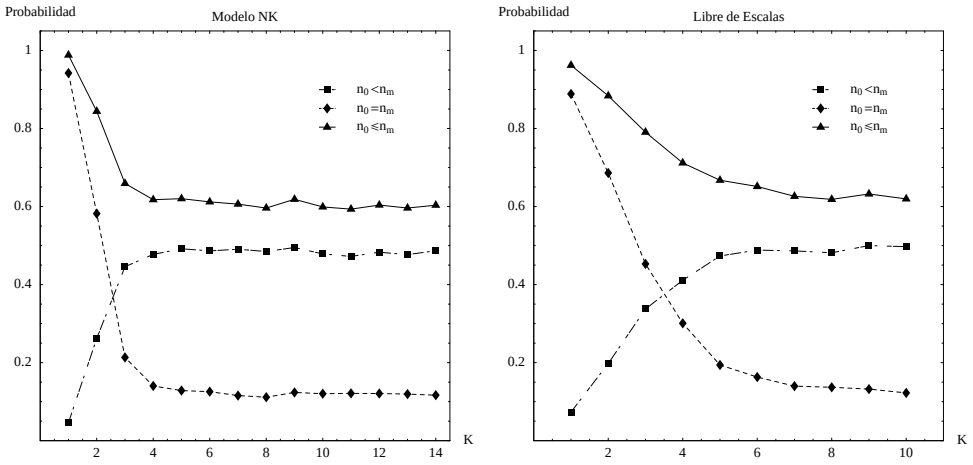

Figura 4.2: Probabilidad de encontrar la misma cantidad o más atractores en la red mutada ($n_o \leq n_m$, triángulos), con n_o y n_m el número de atractores en la red original y en la red mutada respectivamente. Esta probabilidad puede ser descompuesta en la probabilidad de encontrar la misma cantidad de atractores ($n_o = n_m$, rombos) y en la probabilidad de encontrar más atractores ($n_o < n_m$, cuadros). Izquierda: resultados en el *Modelo NK*. Derecha: resultados en redes con topología cercana a la observada en redes reales. El comportamiento general es el mismo para ambas topologías.

de la topología ya que se sabe que en promedio el número de reguladores aumenta conforme aumenta el número de genes en una red [93].

Al comparar las curvas que muestran la probabilidad de tener exactamente el mismo número de atractores o un número estrictamente mayor observamos que una aumenta mientras que la otra disminuye con K, ver Fig 4.2. La probabilidad de tener *exclusivamente más* atractores aumenta con K. Esto refleja el hecho –que explicaremos más adelante– de que las redes mutadas, mientras más caóticas, más tienden a presentar nuevos y más atractores que la red original bajo la perturbación. Sin embargo, tienen la desventaja (las redes mutadas) de perder los atractores originales.

Los atractores de las redes originales pueden sufrir una plétora de transformaciones al mutar la red (añadir un nuevo gen). Para poder comparar los atractores de las redes originales con los atractores transformados he-

Red Original	**Red Mutada**	**Clasificación**
Atractor O_1	**Atractor** M_1	
010100011011011	0101000110110111	
000111000101110	0001110001011101	Identidad
110011001101010	1100110011010100	
Atractor O_2	**Atractor** M_2	
101010111011101	1010101110111011	
010101101101011	0101011011010110	
	*0001000001000101	Expansión
110001011010001	1100010110100011	
	*1000001100110100	
Atractor O_3	**Atractor** M_3	
000001100010011	0000011000100110	
000111011010010	0001110110100100	
*111110011011101		Contracción
111111011011101	1111110110111011	
*111001100101101		
	Atractor M_4	
	*0100000100110010	
	*0010000011010111	Innovación
	*1000010011010010	
	*1001010011010000	

Cuadro 4.2: El ejemplo aquí mostrado proviene de una red original con 15 genes, $K = 2$ y $p = 1/2$. Añadimos un gen al simular duplicación-divergencia –el último en las configuraciones de los atractores de la red mutada y en negrita– que ignoramos al comparar. Los atractore en la red original son O_1, O_2, O_3 y en la mutada M_1, M_2, M_3, M_4. Podemos observar tres transformaciones: identidad, $O_1 = M_1$; expansión, $O_2 \subset M_2$; contracción, $O_3 \supset M_3$. Además hay una innovación, M_4. Las configuraciones marcadas con un asterisco difieren en ambos conjuntos de atractores.

mos ignorado el estado del gen que ha sufrido la duplicación-divergencia, de esta forma comparamos el estado de los genes de la red original contra su estado en el contexto del gen duplicado-divergido. El Cuadro 4.2 muestra resultados típicos al realizar un evento de duplicación-divergencia. En este caso la red original codifica tres atractores O_1, O_2 y O_3; mientras que la red mutada codifica cuatro M_1, M_2, M_3 y M_4. Del Cuadro podemos observar las siguientes transformaciones:

- *Identidad.* Un atractor de la red mutada es idéntico a un atractor de la red original ($O_1 = M_1$).
- *Expansión.* Un atractor en la red mutada contiene todas las configuraciones de un atractor de la red original más algunas configuraciones extra ($O_2 \subset M_2$).
- *Contracción.* Un atractor en la red mutada contiene sólo una parte de las configuraciones de un atractor original ($O_3 \supset M_3$).
- *Innovación.* Un atractor sin contraparte en el conjunto original; un atractor nuevo (M_4).

Debemos tener en mente que además de los cambios en los atractores, la duplicación-divergencia provoca cambios en la cuenca de atracción. Es decir, la perturbación no sólo puede modificar los estados celulares *estables* sino también las rutas para llegar a ellos; discutiremos este punto más adelante. De las transformaciones anteriores sólo consideramos como *biológicamente relevantes* a las identidades y a las innovaciones. A las primeras les llamaremos *conservaciones* por conservar la información previamente codificada en la red. Complementariamente, las innovaciónes expanden la información que codifica la red. Desde el punto de vista biológico, una innovación significa la emergencia (estabilización) de nuevos estados funcionales/fenotipos que permitirían a la célula probar nuevas estrategias.

4.3.2. Resultados para redes homogéneas

Cuantificamos la robustez calculando la probabilidad $P(q)$ de que un porcentaje q de los atractores originales se conserve en la red mutada. La Fig. 4.3 muestra los resultados desde $K = 1$ hasta $K = 4$. Notemos que en $K = 1$ (redes oredenadas) y $K = 2$ (redes críticas) los máximos se encuentran en $q = 0\,\%$ (ningun atractor conservado en la red mutada) y en $q = 100\,\%$ (todos los atractores conservados). Conforme se entra a la región caótica,

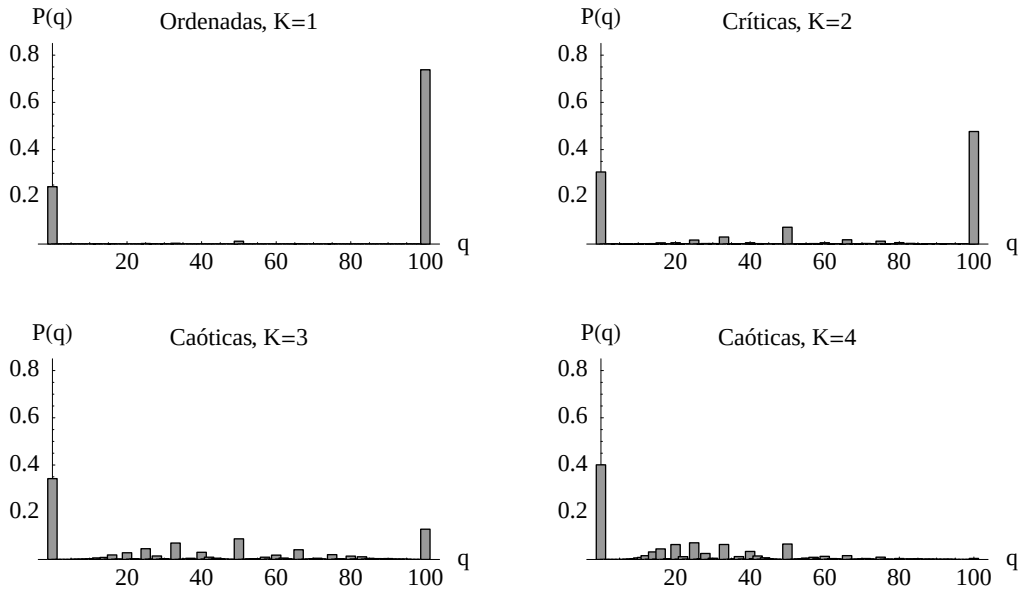

Figura 4.3: Probabilidad $P(q)$ de que un porcentaje q de atractores originales se conserve en la red mutada. Mostramos $P(q)$ para redes ordenadas ($K = 1$), críticas ($K = 2$) y caóticas ($K \geq 3$). En redes críticas y ordenadas el máximo en 100 % indica que, con gran probabilidad, los atractores codificados en la red original se conservan en la mutada. Esta propiedad se pierde al entrar a la región caótica.

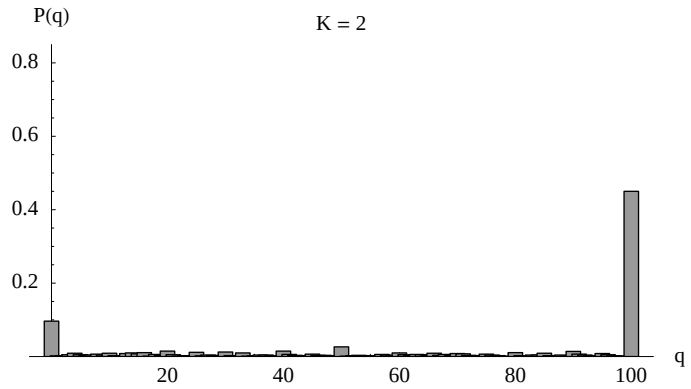

Figura 4.4: Probabilidad $P(q)$ en redes críticas de que un porcentaje q de atractores originales se conserve con la restricción de que cada red que participa en la estadística codifique por lo menos 20 atractores. El máximo que aparece en 0 % en la Fig. 4.3 ha desaperecido lo que indica que éste era producto de redes con pocos atractores que los perdían completamente después de la duplicación-divergencia. Notemos que el máximo en 100 % se conserva.

el máximo en 100 % desaparece y sólo se conserva alguna o ninguna fracción de los atractores originales. Recalquemos que la propiedad de las redes ordenadas y críticas de conservar todos los atractores se da en el caso de divergencia extrema que equivale a añadir un gen con su función y regulaciones completamente al azar.

Los dos máximo que se observan en $P(q)$ podrían deberse a que, en la gran mayoría de las veces, se crean redes con pocos atractores. Por ejemplo, en la gráfica correspondiente a $K = 1$, las redes con dos atractores contribuyen al pico en 50 % (sólo uno de los dos atractores se conserva) o a los picos en 100 %(los dos se conservan) o en 0 %(ninguno se conserva). Redes con un atractor contribuyen exclusivamente a los picos en 0 % y en 100 %. Para descartar que el máximo que observamos en 100 % es debido a este fenómeno hemos calculado para redes críticas[5] ($K = 2$) la misma probabilidad pero con la restricción de que éstas codifiquen veinte o más atractores, Fig. 4.4. Como podemos observar, el máximo en 0 % se ha perdido cediendo su fracción a porcentajes distintos de 0 %. Lo anterior indica que la contribución

[5]Se obtiene el mismo resultado en redes caóticas y oredenadas: el máximo en 0 % desaparece.

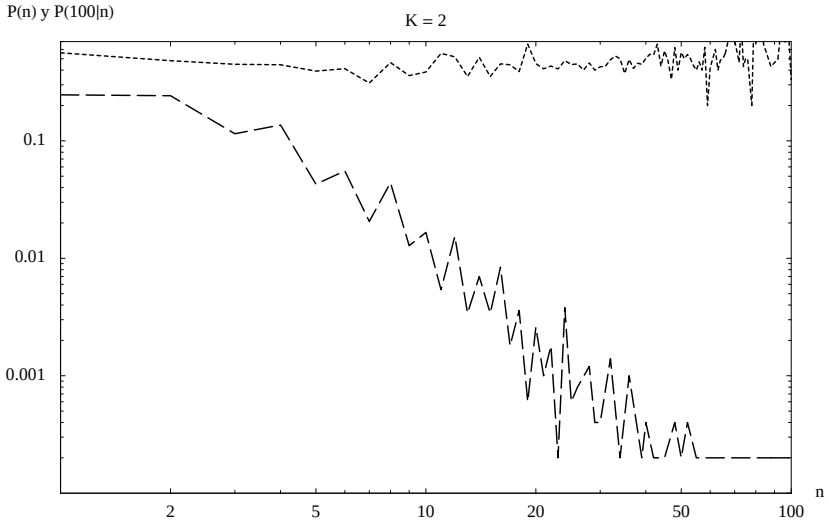

Figura 4.5: Curva inferior: Probabilidad $P(n)$ con la que aperecen redes que codifican n atractores. La distribución es de cola larga e indica que no es raro encontrar redes con un número arbitrario de atractores. Curva superior: Probabilidad $P(100|n)$ de conservar todos los atractores dado que la red original codifica n. Aunque $P(100|n)$ fluctua, su valor es casi constante. Estas gráficas indican que la propiedad de una red de conservar el 100 % de sus atractores es independiente del número de atractores y que su contribución al máximo en 100 % es proporcional a la frecuencia de aparición del número de atractores.

al 0 % en las gráficas de la Fig 4.3 es debida, principalmente, a la perdida de todos los atractores de redes que codifican sólo uno o dos de ellos. El máximo en 100 % permanece.

Gracias a la gráfica anterior, sabemos que el máximo en 100 % no es efecto exclusivo de la conservación de redes con un sólo atractor. Para conocer más en detalle cómo contribuyen las redes que codifican n atractores al máximo en 100 % hemos calculado con que frecuencia $P(n)$ aparecen redes con n atractores y qué probabilidad hay de conservarlos a todos $P(100|n)$, Fig. 4.5. Como podemos ver, la distribución de atractores tiene una cola larga indicando que la red codifica más de un atractor con alta probabilidad. Por otra parte, es interesante observar que la probabilidad $P(100|n)$, pese a las fluctuaciones, es casi constante. Esto es indicativo de que la propiedad de

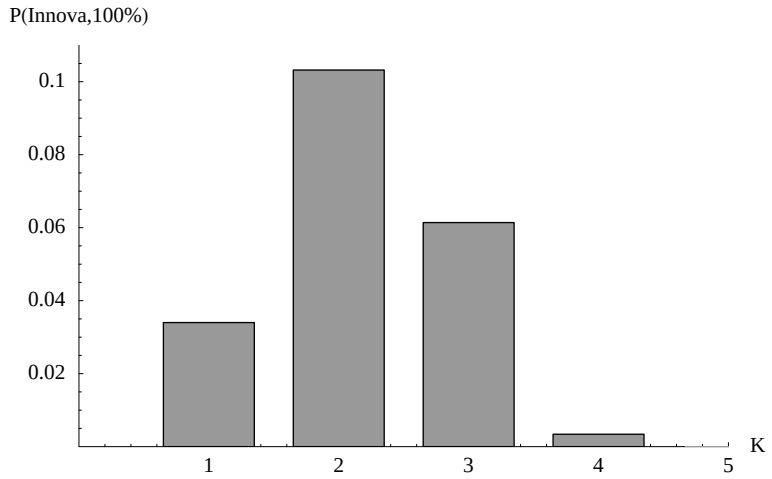

Figura 4.6: Probabilidad de conservar todos los atractores y encontrar alguna innovación después del proceso de duplicación-divergencia como función de K. La gráfica muestra que son las redes alrededor del estado crítico las que poseen la máxima robustez y evolucionabilidad.

conservar todos los atractores es independiente de qué tantos codifique la red en cuestión. Usando estos resultados, llegamos a la conclusión de que, si bien muchas de las conservaciones que contribuyen al máximo en 100 % son debidas a la conservación de las redes que codifican pocos atractores, el que una red codifique varios no es impedimento para que constribuya proporcionalmente al máximo, *i.e.* que la menor contribución de las redes que codifican varios atractores al pico en 100 % no es debida a que los atractores de éstas se conserven con menor probabilidad sino a que aparecen menos redes con varios atractores.

Particularmente importante es la emergencia de nuevos atractores ya que son estos la fuente de nuevos estados funcionales celulares. Los eventos más importantes de innovación desde el punto de vista biológico, suceden cuando aparecen nuevos atractores y además se conservan todos los originales. En la Fig. 4.6 hemos graficado, como función de la conectividad K de la red, la probabilidad $P(innova, 100\,\%)$ de conservar todos los atractores originales y encontrar alguna innovación. El máximo de $P(innova, 100\,\%)$ está situado alrededor de las redes críticas ($K = 2$) indicando que su robustez (conservación total de atractores) y evolucionabilidad (aparición de innovaciones)

es máxima.

4.3.3. Resultados para redes Libres de Escala

Hasta el momento todas nuestras exploraciones las hemos realizado en redes con topología homogénea (a excepción de la gráfica en la Fig. 4.2 con resultados de redes con topología Libre de Escala). Sin embargo, ésta está lejos de ser la más cercana a la observada en las redes de regulación reales. En las tres redes más grandes hasta el momento caracterizadas (red de regulación de *E. coli*, *B. subtilis* y *S. cerevisiae*) se observa que cada gen es regulado en promedio por dos FT ($K = 2$) y que, para una conectividad de salida arbitraria, existe una cantidad finita de FT con esa conectividad. Estudios previos muestran que la conectividad de salida es Libre de Escalas y la de entrada Poissoniana o exponencial [21, 12], ver Apéndice B.

Para generar redes con distribuciones próximas a las observadas, asignamos la conectividad de salida de un gen i en la red usando una variable aleatoria k_s con distribución Libre de Escala. Después, escogemos con probabilidad uniforme k_s genes para ser regulados por el mismo gen i. Haciendo esto para cada gen en la red, la distribución de la conectividad de entrada será Poissoniana.

Comenzamos con una red original de 20 genes para perturbarla mediante el proceso de duplicación-divergencia y llegar a una red mutada con 21 genes. Nuevamente, la divergencia es extrema por lo que el nuevo gen no tiene ninguna relación con algún otro en la red. Cuando añadimos el gen duplicado-divergido seleccionamos su conectividad de salida de una distribución Libre de Escalas. De esta forma, tratamos de conservar en la red mutada la distribución que se observa en la realidad. La selección de la conectividad de salida concuerda con el hecho de que, hasta donde se sabe, la tasa de duplicación-divergencia no depende de la conectividad del gen [94, 95, 96].

En la Fig. 4.7 mostramos la probabilidad $P(q)$ de conservar una fracción q de los atractores originales. Las distintas gráficas en la figura corresponden a distintos valores del exponente γ en la distribución Libre de Escala. Cada valor del exponente es tal que el número de reguladores promedio K va de uno a cuatro[6]. En general los resultados son similares a lo que se obtiene

[6]El número de reguladores promedio K en una red con topología Libre de Escala, n

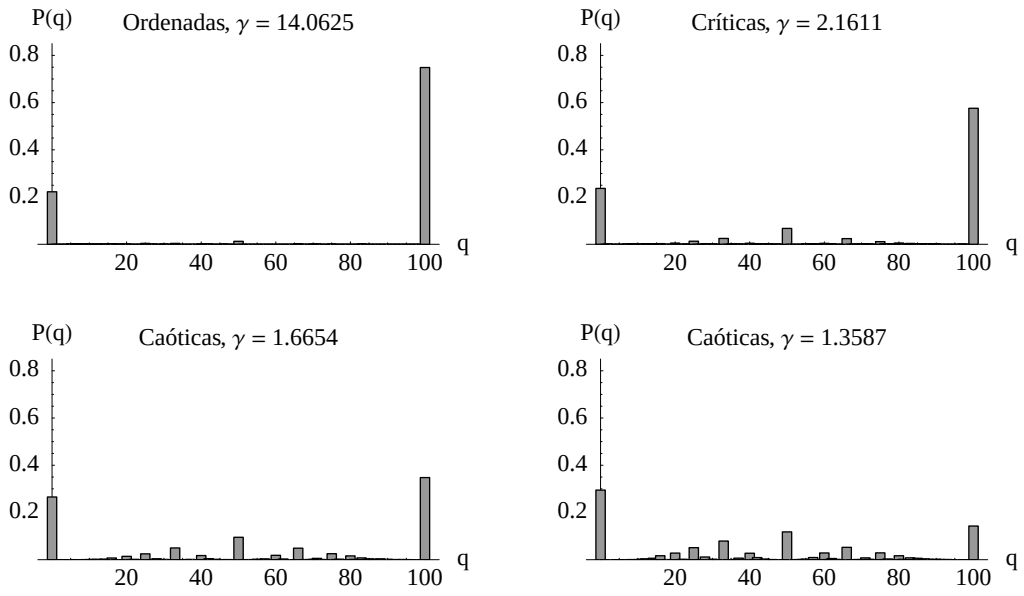

Figura 4.7: Probabilidad $P(q)$ de conservar una fracción q del total de los atractores en la red original. El resultado es para redes con topología Libre de Escalas en la conectividad de salida y Poissoniana en la de entrada. Cada gráfica está etiquetada con el valor del exponente γ de la distribución Libre de Escalas. El número de reguladores promedio es $K = 1$ (γ = 14,0625), $K = 2$ (γ = 2,1611), $K = 3$ y $K = 4$ (γ = 1,6654 y 1,3587). Notemos que la robustez es mayor que la observada en redes con topología homogénea.

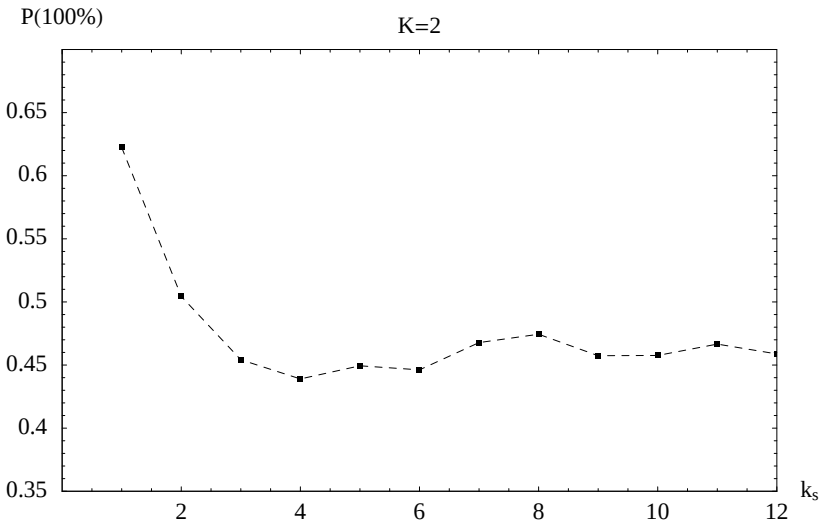

Figura 4.8: Probabilidad de conservar todos los atractores $P(100\,\%)$ como función del número de genes que regula el gen duplicado-divergido k_s. La gráfica evidencia que seleccionando con mayor probabilidad una conectividad de salida pequeã (*e.g.* $k_s = 1$, $k_s = 2$) –como lo favorece una Ley de Potencias– la robustez aumenta.

para redes con topología homogénea: existen dos máximos locales en $q = 0\,\%$ y $q = 100\,\%$ y la prominencia del pico en $100\,\%$ disminuye conforme transitamos de la región ordenada a la caótica. Sin embargo, existe una diferencia importante: la robustez observada (pico en $100\,\%$) es mayor en las redes con topología Libre de Escala. En $K = 1$ la robustez es mayor por un factor de 1.0144 y en $K = 4$ por 32.4091. Incluso para $K = 4$ los máximos en $0\,\%$ y en $100\,\%$ son comparables.

Para investigar por qué las redes con topología Libre de Escala son más robustas, hemos gráficado la probabilidad de conservar todos los atractores $P(100\,\%)$ contra el valor de la conectividad de salida k_s del gen que sufre duplicación-divergencia en redes críticas ($K = 2$), Fig. 4.8. Es decir, hemos perturbado sistematicamente a la red agregando un nuevo gen con grado de salida uno, dos, tres, etc. y calculado la probabilidad de conservar todos los atractores. La probabilidad $P(100\,\%)$ disminuye rápidamente conforme

genes y exponente γ es $K = (\sum_{k=1}^{n} k^{1-\gamma})/(\sum_{k=1}^{n} k^{-\gamma})$. El exponente γ puede ser ajustado para obtener un valor particular de K.

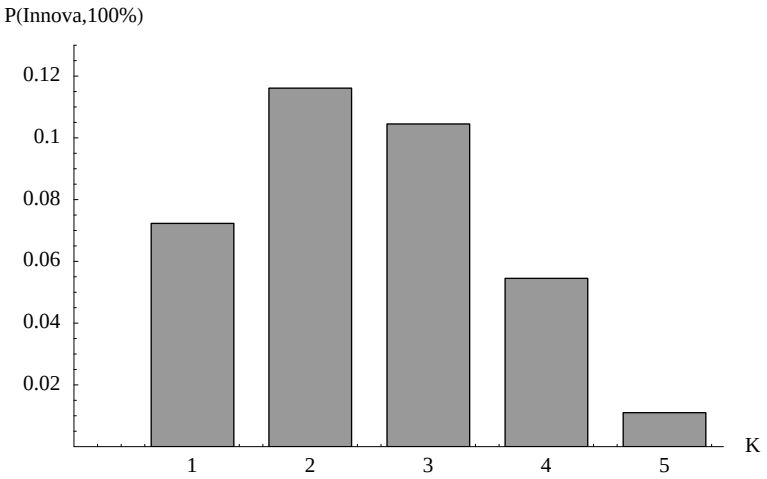

Figura 4.9: Probabilidad de conservar todos los atractores y encontrar alguna innovación después del proceso de duplicación-divergencia como función del número de reguladores promedio K. Nuevamente, las redes críticas maximizan la robustez y evolucionabilidad.

aumenta k_s aunque no tiende a cero. Refiriéndonos al valor de $P(100\,\%)$ para redes críticas en la Fig. 4.7 observamos que su valor es próximo a 0,6; valor que casi coincide con la robustez obtenida en la Fig. 4.8 cuando el gen añadido regula a un sólo gen ($k_s = 1$). Por lo tanto, el aumento de robustez proviene de que los genes duplicados-divergidos que más conservan los atractores originales (genes con grado de salida $k_s = 1$) son también los que más se escogen en una simulación típica. Notemos que éste aumento de robustez es *gratuito* ya que las constricciones intrínsecas del mecanismo de duplicación-divergencia en un contexto de red con topología Libre de Escalas, automáticamente favorece una álta conservación. En este caso, la constricción intrínseca consiste en que las duplicaciones de reguladores globales –que están asociados a una robustez baja– son un evento raro.

Al igual que las redes con topología homogénea, las redes Ley de Potencias conservan e innovan de forma máxima cuando el número promedio de reguladores K se encuentra alrededor de dos, es decir cuando las redes están en la región crítica, ver Fig. 4.9. Sin embrago, notemos que el máximo es menos pronunciado que en el *Modelo NK*, ver Fig. 4.6.

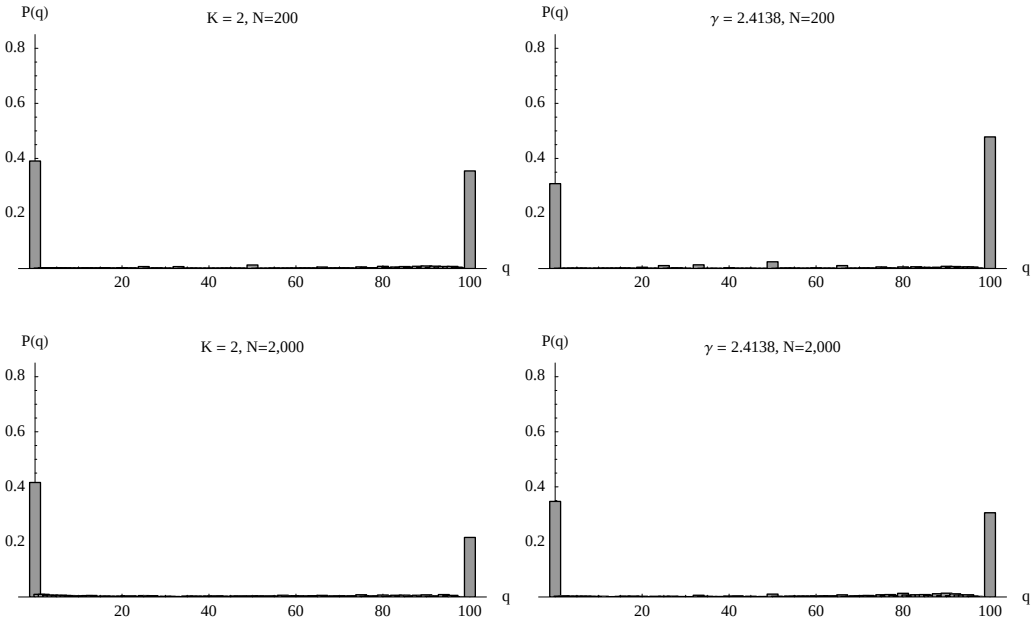

Figura 4.10: Probabilidad $P(q)$ de conservar una fracción q de atractores originales en la red mutada. Los resultados son para redes con 200 (gráficas superiores) y 2,000 (gráficas inferiores) genes en la región crítica, con topologías homogénea y Ley de Potencias a la izquierda y derecha respectivamente. Aunque el tamaño de la red ha aumentado uno y dos ordenes de magnitud, los resultados son casi idénticos a los de redes con 20 genes. Los máximos en 100 % son menores que sus equivalentes en redes de 20 genes debido, probablemente, a que la estrategia de búsqueda de atractores no detecta aquellos con cuencas chicas. Calculamos las gráficas usando los resultados de 20,000 redes.

4.3.4. Redes Grandes

Debemos asegurarnos que los resultados hasta ahora encontrados no dependen fuertemente de la cantidad de genes que conforman a la red. Para estudiar el efecto del tamaño en la robustez de la red, ampliamos en uno y dos ordenes de mágnitud el número de genes, es decir caracterizamos redes con 200 y 2,000 elementos.

En las secciones pasadas, el estudio de redes de 20 genes nos permitía explorar completamente el espacio de configuraciones y encontrar todos los atractores. Sin embargo, extender esta estrategia a redes significativamente

más grande, es computacionalmente prohibitivo. Como el espacio de configuraciones crece exponencialmente con el número de genes en la red, una exploración exhaustiva para detectar atractores, configuración por configuración, es imposible. Con sólo un aumento de 10 genes a una red que ya contiene 20, los tiempos de cálculo y la memoria requerida se tienen que multiplicar, por lo menos, por mil. La estrategia para poder caracterizar redes un orden de magnitud más grandes consiste en explorar sólo un subconjunto de todas las configuraciones posibles y encontrar los atractores a los que éstas llevan. La estrategia tiene como desventaja que algunos atractores no serán detectados, particularmente aquellos con una pequeña cuenca de atracción.

En redes con 200 y 2,000 genes, implementamos una estrategia de busqueda para poder detectar hasta 500 atractores, cada uno de ellos hasta de 1,000 configuraciones. El tamaño del espacio de configuraciones que exploramos es de un millón –fracción ínfima de todo el espacio (aprox. 10^{60} configuraciones). Debido al costo computacional, caracterizamos sólo redes críticas que, cómo mostramos en el capítulo anterior, son las más cercanas a las redes reales. La Fig. 4.10 muestra, para redes con topologías homogénea y Libre de Escala, la probabilidad $P(q)$ de conservar una fracción q de todos los atractores encontrados. Recordemos que en las secciones anteriores, q representa la fracción de *todos* los atractores de la red original que son conservados por la mutada. Aquí, q representa la fracción de todos los atractores que se pudieron encontrar en la red mutada y que coinciden con los que se pudieron encontrar en la red original. En las gráficas podemos observar nuevamente la estructura ya obtenida en las redes con 20 genes: existen dos máximos, uno en 0 % y otro en 100 % que indica que las redes críticas, aún un orden de magnitud más grandes, conservan con alta probabilidad los atractores. Existe, sin embargo, una pequeña disminución del pico en 100 % siendo el observado en redes con 20 genes un poco mayor que el observado en redes con 200 y éste a su ves un poco mayor que el observado en redes con 2,000 genes. Pensamos que esto se debe a que no detectamos todos los atractores –posiblemente los de cuenca de atracción más chica– en la red original y en la mutada y no a que el aumento de genes en la red provoque una disminución en su robustez.

Ya hemos mencionado que si las redes sólo tuvieran un atractor, la aparición de los dos máximos no sería tan sorprendente: sólo habría dos opciones, conservación o pérdida. Mostramos que ésto no es el caso graficando la distribución $P(n)$ del número de atractores n y la probabilidad $P(100|n)$ de

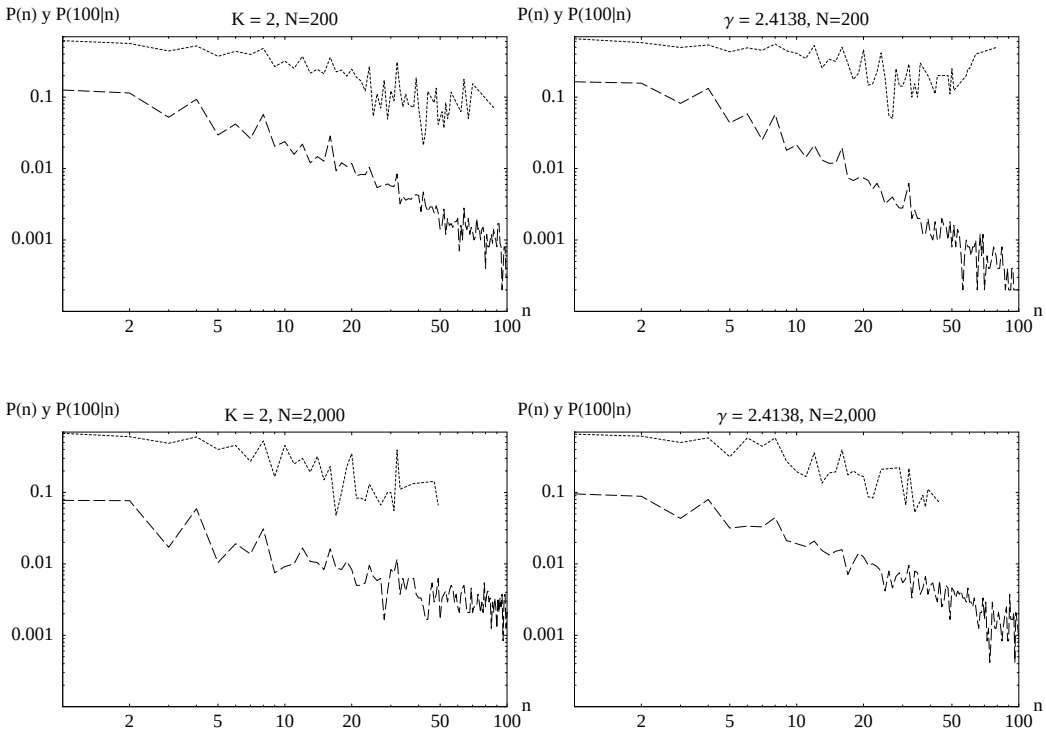

Figura 4.11: *Gráficas superiores:* Resultados para redes con topología homogénea y Libre de Escalas con 200 genes. *Gráficas inferiores:* Redes con topología homogénea y Libre de Escalas con 2,000 genes. En cada gráfica, *curva inferior:* Probabilidad $P(n)$ con la que aparecen redes que codifican n atractores; *curva superior:* Probabilidad $P(100|n)$ de conservar todos los atractores dado que la red original codifica n. Las curvas dan indicios del resultado contraintuitivo de que conservar todos los atractores en una red es casi independiente del número que ésta codifique. Los resultados fueron obtenidos de 20,000 redes para cada topología y cada tamaño de red en la fase crítica.

conservar el 100 % dado que hay n en la red original; ésto lo hemos calculado para redes homogéneas y Libre de Escalas con 200 y 2,000 genes, Fig. 4.11. Notemos primero que independientemente de la topología, la probabilidad $P(n)$ es de cola larga. Esto indica que, con alta probabilidad, la red original codifica más de un atractor. Por otra parte, $P(100|n)$, a comparación de $P(n)$, decrece lentamente al aumentar n. Lo anterior indica que no es gran impedimento el que una red tenga varios atractores para que todos se conserven. De hecho, es probable que la ligera caida que detectamos en $P(100|n)$ (un poco más pronunciada en redes con 2,000 genes) se deba a que nuestra estrategia de búsqueda pierde algunos atractores. De esta forma los resultados serían completemente análogos a los obtenidos en la Fig. 4.5.

Es posible que los resultados obtenidos para redes con 200 y 2,000 genes puedan ser extendidos a redes con 20,000 genes. Sin embargo, ir más allá podría no ser biológicamente relevante ya que en las redes de regulación sólo una pequeña parte se encarga del control. Como vimos en el capítulo anterior, lo que se conoce de las redes de regulación reales muestra que menos de un 10 % calcula la respuesta de toda la red, *i.e.* qué genes efectores se encuentran en qué condiciones. Por poner un ejemplo, el número de genes de *E. coli* en esta fracción que calcula la respuesta (cabeza de la red), es de apenas 103 genes de 1,328 que contiene la red.

4.3.5. El paisaje de atractores

Mostramos esquemáticamente la robustez y evolucionabilidad de las redes de regulación en la Fig. 4.12 para una red con 15 genes en la región crítica. En ella se encuentra el paisaje original de atractores y el paisaje transformado por la adición de un nuevo gen. Los atractores originales O_1, O_2 y O_3 son conservados en los atractores M_1, M_2 y M_3 del paisaje transformado. Además de las conservaciones, surgen innovaciones M_4, M_5 y M_6. Vemos que la dinámica de la red es robusta al conservar los atractores originales y además es evolucionable ya que el paisaje transformado presenta nuevos. Esto es un caso típico en redes que se encuentran en la fase crítica.

Cómo ya hemos mencionado, el que se conserven los atractores originales no implica que el resto del paisaje no se reconfigure; en la Fig. 4.12 las cuencas de los distintos atractores originales sufren cambios después de la adición del nuevo gen. Para ilustrar de una mejor manera este fenómeno, marcamos de cierto color a todas las configuraciones pertenecientes a una

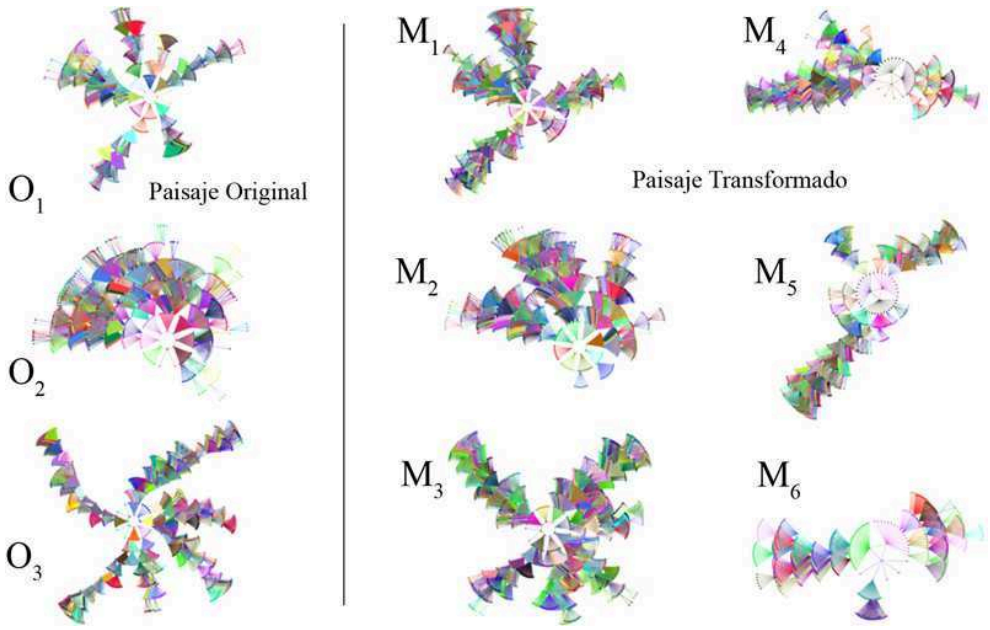

Figura 4.12: Paisajes de atractores original y transformado. El paisaje original tiene tres atractores con sus respectivas cuencas, O_1, O_2 y O_3. Estos atractores son conservados en la red mutada y están representados en el paisaje transformado por M_1, M_2 y M_3; los subíndices entre atractores originales y mutados se corresponden. Además, en el paisaje transformado, aparecen tres innovaciones M_4, M_5, M_6. Notemos que las cuencas de los atractores conservados han sufrido cambios.

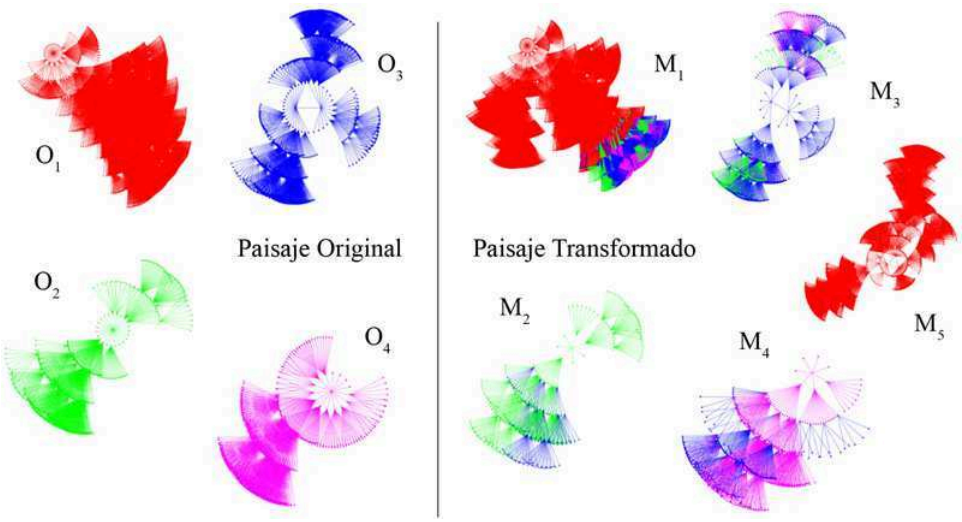

Figura 4.13: Ilustración de la transformación de las cuencas. En el paisaje original, marcamos las configuraciones de cada atarctor y su cuenca con cierto color. Respetando éste código de color en el paisaje transformado, podemos saber qué configuraciones cambiaron de cuenca.

sóla cuenca de atracción del paisaje original, hacemos ésto para cada una de las cuencas, Fig. 4.13. Este mismo código de color lo respetamos al construir el paisaje transformado. Podemos notar que, aunque en el paisaje transformado las cuencas de atracción se han conservado en gran medida, existen configuraciones que antes no pertenecían a las mismas, es decir, las cuencas transformadas son de varios colores.

Biológicamente, la reconfiguración de las cuencas de atracción indica que las vías de diferenciación pueden cambiar aún cuando la perturbación duplicación-divergencia deje invariante el conjunto de atractores originales. Esto ya ha sido señalado en el modelo continuo de la red encargada de la segmentación polar en la mosca *D. melanogaster* [97]. En la red se ha realizado el experimento *in silico* de reconectar a los genes o duplicarlos, dando como resultado el cambio de la evolución temporal de los niveles de expresión (reconfiguración de la cuenca) pero conservando el patrón correcto de la expresión en el estado estacionario de la dinámica (atractor conservado).

4.4. Conclusión.

Robustez y Evolucionabilidad, Características Esenciales de la Red de Regulación

El grado de éxito de un modelo radica en qué tan bien sintetiza las características esenciales del fenómeno al que se refiere. Como ya hemos mencionado, Kauffman en su modelo, impone lo que él cree son las constricciones intrínsecas esenciales a la regulación. Esto es, que un gen que regula puede a su vez ser regulado; característica que, en un conjunto de genes, forma una red de regulación. Lo anterior provoca que sólo un pequeño conjunto de configuraciones sea estable, que existan fases del sistema dinámico de regulación con caracteristicas particulares, etc. En años recientes, la confianza que se tiene sobre estas constricciones intrínsecas esenciales y lo que implican ha aumentado gracias al hecho, comprobable experimentalmente, de que los estados estables de redes construidas con el conocimiento biológico actual se corresponden con los patrones de expresión observados [26, 25, 32, 33]. Además, distintos hechos experimentales (directos e indirectos) que se van acumulando ([70, 71, 72], ver también Cap. 3 de éste mismo trabajo) muestra que las redes reales se encuentran en verdad proximas a la fase crítica. Esto permite situar en un marco teórico preciso el conocimiento empírico que indica que los organismos son robustos y adaptables.

En la medida en que hechos similares a los expuestos en el párrafo anterior sigan apuntalando la confianza en el modelo podemos decir que las mismas constricciones intrínsecas dan lugar a que una red de regulación sea robusta y evolucionable. Esto quiere decir que la propia definición del modelo facilita que configuraciones antes inestables en cuencas de atracción se estabilicen dando origen a nuevos atractores sin que ello conlleve la perdida de los ya presentes. Esto constituiría un ejemplo más del *orden gratuito*[7] que es complementario al orden que se alcanza a través de selección natural.

En la misma vena (suponiendo que el modelo de Kauffman captura características esenciales de la regulación), los hechos presentados en este capítulo apoyan el punto de vista de que las redes operan en la fase crítica: la robutez y evolucionabilidad máximas en esta fase son compatibles con lo observado en la redes de regulación reales.

Debemos mencionar que, además de la topología homogénea y Libre de Es-

[7]Expresión acuñada por Kauffman.

cala en la conectividad de salida, también estudiamos redes con topología Poissoniana y Libre de Escala en el grado de entrada. Los resultados son cualitativamente idénticos: los atractores se conservan con gran probabilidad en redes ordenadas y críticas y la robustez y evolucionabilidad es máxima alrededor de la fase crítica. Esto nos hace pensar que las constricciones topológicas no son esenciales para que exista robustez y evolucionabilidad bajo duplición-divergencia.

Desde el punto de vista *global* que otorga el modelo de Kauffman, la atención se centra en los patrones de expresión (atractores) los cuales tienen una identidad propia que se identifica con alguna función característica. Esta es la perspectiva complementaria a asociar a cada proteína una función (perspectiva *local*)[8]. Cambiando de perspectiva, global o local, podemos analizar el surgimiento de nuevos estados funcionales celulares. Desde una visión local, para que se cree un nuevo patrón de expresión se tiene que dar toda una serie de eventos (cada uno de ellos respaldado y justificado por la selección natural) para, posiblemente, llegar al nuevo patrón. Desde el punto de vista global, lo que sucede es que la perturbación duplicación-divergencia, *estabiliza* un conjunto de configuraciones antes transitorias creando atractores que podrían ser seleccionados a favor o en contra.

La duplicación-divergencia desde la perspectiva global podría explicar la función paradójica de varias proteínas de señalización como *ras*. Esta proteína se ha identificado como promotora de muerte celular, crecimiento y diferenciación [32]. Una explicación plausible es que, suponiendo que el crecimiento sea un estado funcional previo a los otros dos donde ya se presentaba *ras*, eventos de duplicación-divergencia estabilizaron nuevos patrones de expresión que se convirtieron en apoptosis y cierto nuevo tipo celular, involucrando a *ras* en funciónes contradictorias o distintas. Desde el punto de vista global, los patrones de expresión tienen una identidad propia siendo las proteínas actores que cambian de rol dependiendo del patrón.

[8]El Dr Huang hizo patentes estas distintas visiones en una charla dada en el ciclo de conferencias *Nuevas Direcciones en Redes de Regulación Genética*, Cuernavaca, Morelos, México, 2006

Apéndice A

Microarreglos

Los microarreglos permiten descubrir conjuntos de genes en algún organismo que cambian sustancialmente su expresión en distintas condiciones experimentales. Esto implica tener por lo menos dos experimentos de microarreglo. Comunmente a uno se le designa experimento *control* y al otro, experimento *objetivo*. Antes de tratar de encontrar qué genes cambian su expresión de un experimento a otro es necesario procesar las intensidades crudas que se obtienen del experimento de microarreglo. El procesamiento consiste en la eliminación del ruido de fondo y la normalización.

- *Eliminación del ruido de fondo.* Los microarreglos son una tecnología particularmente ruidosa: la hibridación puede no ser perfecta; mensajeros (aún despúes del lavado del microarreglo) pueden quedar depositados sin estar hibridizados, etc. Esto provoca que las intensidades crudas no reporten de forma fiel la cantidad de ARNm presente en la condición que se examina. Una estimación del ruido se obtiene al registrar la intensidad de areas adyacentes al conjunto de sondas. La lectura de este ruido se hace de forma automatizada y los archivos con las intensidades crudas vienen acompañados, por lo general, del ruido adyacente. Se debe restar el ruido a las intensidades crudas.

- *Normalización.* Sucede que, aún siguiendo el protocolo experimental, lecturas crudas de dos microarreglos de una misma muestra no son comparables: por lo general uno de los dos microarreglos presenta un aumento de expresión en casi todos sus genes. Esto es incorrecto ya que células de una misma muestra expresan en promedio el mismo material genético en las mismas proporciones. El problema se soluciona al normalizar las intensidades de cada microarreglo a una intensidad media

igual a la unidad: se calcula la mediana de expresión del microarreglo y se divide por esta cantidad a todas las intensidades; después de la transformación, la nueva mediana es la unidad[1]. Cuando las muestras a comparar no son las mismas –*i.e.* el control y el experimemto objetivo son distintos– no tiene por que suceder que también sea similar la intensidad media de ambas. Sin embargo, uno no espera cambios radicales en toda la expresión del genoma: en distintas condiciones, la célula debe sobreexpresar o inhibir a algunos genes, dejenado a la gran mayoría sin variación. Usando esta hipótesis, hacemos comparables las lecturas de las dos condiciones nuevamente normalizándolas a una intensidad media igual a la unidad, ver Fig. A.1.

Cuando se ha corregido el ruido y se han hecho comparables los distintos microarreglos, se puede calcular sin más el logaritmo de la razón entre la señal objetivo y la señal control. El logaritmo es base dos para poder identificar fácilmente cuantas veces se ha duplicado o se ha dividido a la mitad la expresión de algún gen entre condiciones.

La normalización y la estimación del ruido son dos problemas cuya solución aún no es satisfactoria. En la actualidad se siguen proponiendo métodos para eliminar el ruido y hacer comparables distintas condiciones [98]. Lo arriba expuesto representa las soluciones más simples y más ampliamente usadas. Los datos de microarreglo usados en este trabajo fueron procesados según acabamos de mostrar cada vez que fue necesario[2].

A.1. Microarreglos Usados en la Inferencia

Relación de los identificadores (ID's) con los que se puede encontrar en las distintas bases de datos a los microarreglos usados en la inferencia.

Bacillus subtilis

[1] Se usa la mediana porque es una mejor estimación del valor típico en una muestra. El valor de la mediana se ve menos afectado que el del promedio cuando la muestra contiene valores muy alejados de la mayoría.

[2] Los datos que usamos en la inferencia provienen de distintas fuentes y no todos tienen el mismo procesamiento. Algunas veces, los datos estaban completamente procesados (ruido eliminado y normalizados) otras veces completamente crudos.

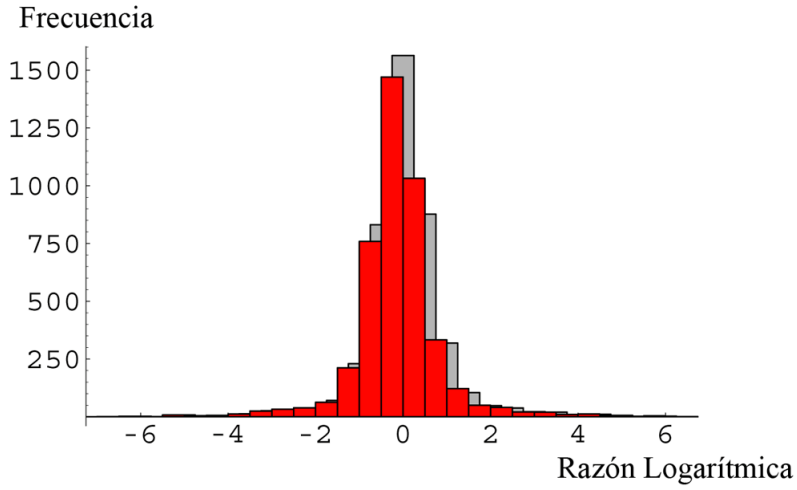

Figura A.1: Histograma de razones logarítmicas entre dos experimentos de microarreglo (tomados de *The Blattner Lab*, `www.genome.wisc.edu/`). Los experimentos comparan, en *E. coli*, los cambios de expresión entre una condición aeróbica y otra anaeróbica. El histograma al frente, es de razones logarítmicas de datos sin normalizar. El histograma al fondo, es de razones logarítmicas de datos normalizados usando la mediana. Notemos que el histograma con datos no normalizados está corrido a la izquierda reflejando una falsa subexpresión de todo el genoma.

Los experimentos los adquirimos de `www.genome.jp/kegg/expression`, 69 en total. Los datos ya se encuentran normalizados, basta con restar el ruido de fondo.

ex0000263	ex0000264	ex0000265	ex0000266	ex0000267	ex0000268	ex0000269
ex0000270	ex0000271	ex0000272	ex0000273	ex0000274	ex0000275	ex0000276
ex0000277	ex0000278	ex0000279	ex0000280	ex0000281	ex0000282	ex0000283
ex0000284	ex0000285	ex0000286	ex0000260	ex0000261	ex0000262	ex0000259
ex0000258	ex0000358	ex0000369	ex0000370	ex0001750	ex0001751	ex0001752
ex0001753	ex0000824	ex0001597	ex0001598	ex0001599	ex0001600	ex0001437
ex0001439	ex0001440	ex0000798	ex0000360	ex0000940	ex0000941	ex0000942
ex0000943	ex0000944	ex0000945	ex0000340	ex0000377	ex0000782	ex0000381
ex0001360	ex0000659	ex0000660	ex0000661	ex0000744	ex0000746	ex0000745
ex0000747	ex0000749	ex0000758	ex0000748	ex0000785	ex0000395	

Escherichia coli

Los experimentos pueden encontrarse en *The Stanford Microarray Database*, `http://genome-www5.stanford.edu`, 107 en total. Los datos fueron obtenidos normalizados, sin ruido y como razones logarítmicas base 2. Intensidades marcadas como no recomendables, fueron eliminadas.

ID del experimentador: **JONB**

8377	8379	8536	25831	15343	13838	15341
15342	13840	15337	15338	15336		

ID del experimentador: **KHODURSK**

5265	5277	5278	5268	5281	5272	5273
1642	1646	1644	1647	1649	1643	1650
1285	1290	1292	1911	1908	1912	1913
1914	1909	1915	1916	14832	14830	14831
14829	1596	2353				

ID del experimentador: **KANGSEOK**

32748	32749	32746	32757	32759	32760

ID del experimentador: **CHRISM**

13076 13077 36049 36051 36054

ID del experimentador: **MBSUE**

19887 19888 19627 19628 19629 18991 18989
19343 19344 19345 18990 19886 19807 19810
19812

Otros 45 experimentos se obtuvieron al solicitarlos directamente a los autores de [68, 69].

Saccharomyces cerevisiae

Los experimentos pueden encontrarse en *The Stanford Microarray Database*, `http://genome-www5.stanford.edu`, 223 en total. Los datos fueron obtenidos normalizados, sin ruido y como razones logarítmicas base 2. Intensidades marcadas como no recomendables, fueron eliminadas.

ID del experimentador: **AGASCH**

7529	7530	4779	691	4778	12801	4780
4781	4782	7528	692	989	990	991
992	985	986	6377	987	988	993
6679	6681	6685	6688	6668	6672	6676
972	975	963	964	961	962	1661
974	971	2558	2556	2064	2557	2554
2555	830	831	834	836	5357	838
840	842	5245	5243	5246	5244	5239
5237	5240	5242	984	1250	1257	1251
1252	1253	1249	1254	1255	977	983
978	979	980	976	981	982	8525
6364	6439	8522	6362	8523	8524	8520
6349	6351	6344	6357	6358	6354	8521
8528	8526	825	1108	1106	824	1155
819	1139	1104	827	815	817	812
5358	4872	4874	7547	7549	7535	7536
7537	7538	7539	4785	4787	4786	2560
2559	4784	4783	4789	809	814	1258
6361	811	810	813	7540	4047	4048
7541	7542	877	878	5187		

ID del experimentador: **SPELLMAN**

686	684	681	680	1686	685	512
513	514	515	516	517	518	519
8227	8228	8229	8230	8233	8234	8235
584	585	597	598	599	8193	8217
8195	8197	8204	8210	8213	8247	497
781	797	799	800	801	802	106
107	108	12744	109	115	410	409
408	407	406	2144	116	117	118

ID del experimentador: **JDERISI**

1303	1311	1312	1302	1309	1313	1310
8282	8284	8287	8286	8291	7998	7999
8000	8003	7893	7895	6854	6859	6824
6841	6826	6856	6834	6836	6839	7985
8289						

Apéndice B

Topología de las Redes de Regulación

La importancia central de las redes de regulación impulsa en gran medida el desarrollo de técnicas experimentales y computacionales para desvelarlas. Actualmente, se pueden distinguir cuatro formas para construir redes: por curación [13, 14, 12], por ortología [91, 92], por experimentos ChIp-chip [99] y por inferencia estadística[100, 101]. La curación depende del conocimiento de expretos que revisan evidencia experimental publicada en artículos arbitrados para establecer y validar regulaciones entre los genes. La ortología, básicamente, busca genes reguladores y regulados homólogos en organismos cercanos y propone que, para homólogos reguladores-regulados, existen las mismas regulaciones. Los experimentos ChIp-chip detectan, para un FT especialmente marcado, todos los sitios de unión a cromatina (regiones reguladoras). Cada sitio de unión establece una regulación entre el FT y el gen al que pertenece la región reguladora. Por último, la inferencia estadística, usando experimentos de microarreglo, construye redes buscando cambios correlacionados entre genes [101] o proponiendo las redes que mejor expliquen los cambios observados [100]. La construcción de las redes por curación es la referencia a la que deben regresar las tres últimas técnicas para validar sus resultados. A su ves, las tres últimas técnicas reducen mucho el conjunto de genes entre los que se deben buscar regulaciones. Esto crea un cíclo *propuesta de regulación - validación* que en principio podría llegar a ser muy eficiente.

Las redes de *B. subtilis* [14], *E. coli* [13], *A. thaliana* [26, 27] y *D. melanogaster* [25] aquí usadas fueron construidas por curación. La red de *S.*

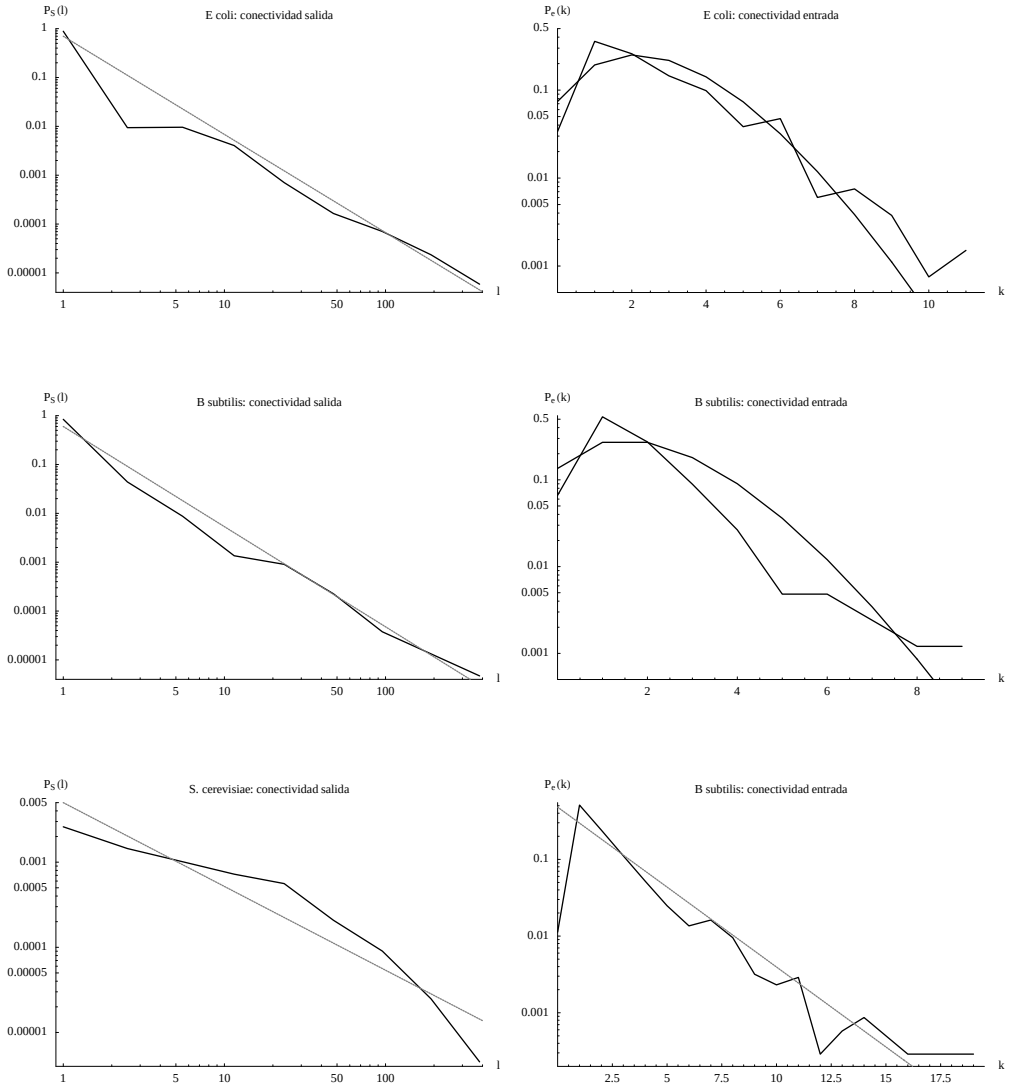

Figura B.1: Distribuciones de conectividad de salida $P_s(l)$ (izquierda) y de entrada $P_e(k)$ (derecha) de tres microorganismos: *E. coli*, *B. subtilis* y *S. cerevisiae*. Datos experimentales línea negra; ajuste línea gris. Una distribución Libre de Escalas aproxima bien a $P_s(l)$ en los tres microorganismos. El valor del exponente es: γ =2.01 para *E. coli*, γ =2.05 para *B. subtilis* y γ = 0.984 para *S. cerevisiae*. $P_e(k)$ es bien aproximada por una Poissoniana en *B. subtilis* y *E. coli* con promedio K =1.8 y K =2.6 respectivamente. $P_e(k)$ en *S. cerevisiae* es más próxima a uan distribución exponencial $\alpha e^{-\alpha k}$ con α =0.5.

cerevisiae es la unión de una red obtenida por curación [12] y otra obtenida por experimentos ChIp-chip [99]. La unión de las redes fue llevada a cabo por Luscombe *et al.* [11]. Damos una representación gráfica de cada una de las redes de regulación de los microorganismos en tres figuras: *E. coli* Fig. B.2, *B. subtilis* Fig. B.3 y *S. cerevisiae* Fig. B.4.

B.1. Distribucion de Entrada $P_e(k)$ y de Salida $P_s(l)$

Para implementar el modelo de Kauffman se pueden usar distribuciones arbitrarias en la conectividad de entrada $P_e(k)$ y de salida $P_s(l)$. Sin embargo, las redes de regulación siguen distribuciones específicas. En la Fig. B.1 mostramos las distribuciones $P_e(k)$ y $P_s(l)$ de los tres microorganismos. Vemos que para *E. coli*, $P_e(k)$ es correctamente aproximada por una distribución de Poisson con promedio K =2.6 y que $P_s(l)$ es bien aproximada por una distribucion Libre de Escala con exponente $\gamma = 2.01$. Análogamente, la distribucion de entrada de *B. subtilis* es bien aproximada por una distribución de Poisson con promedio K =1.8 y la de salida por una Libre de Escala con exponente $\gamma = 2.05$. Finalmente, la distribución de entrada de *S. cerevisiae* es más proxima a una distribución exponencial $\alpha e^{-\alpha k}$ con α =0.5 y la de salida a una Libre de Escala con exponente γ =0.984. Independientemente de los ajustes, los resultados muestran que la connectividad de entrada es de cola corta y la de salida de cola larga. Esto nos permite hablar de un número de reguladores promedio y de la existencia de reguladores globales.

B.2. Estructura Jerárquica

Conocer la topología de la red no nos dice nada sobre que tan "enredada" está la misma, es decir, si se puede o no establecer una estructura de control jerárquica. En el Cap. 3 vimos que gran parte de la red esta organizada de forma jerárquica y controlada por una cabeza constituida por menos del 8 % de los genes totales. Para tener una idea de que tán jerárquica es a su vez la cabeza hemos ordenado a ésta, para el caso de *E. coli*, de la siguiente manera (ver Fig. B.5): separamos a la red en capas de tal forma que las superiores pueden ejercer control sobre las inferiores o *sobre sí mismas* a través de cíclos de regulación o asas de retroalimentación. Los genes en la mitad izquierda de cada capa son regulados por al menos un gen en alguna capa superior. Los genes a la derecha tienen la particularidad de no ser

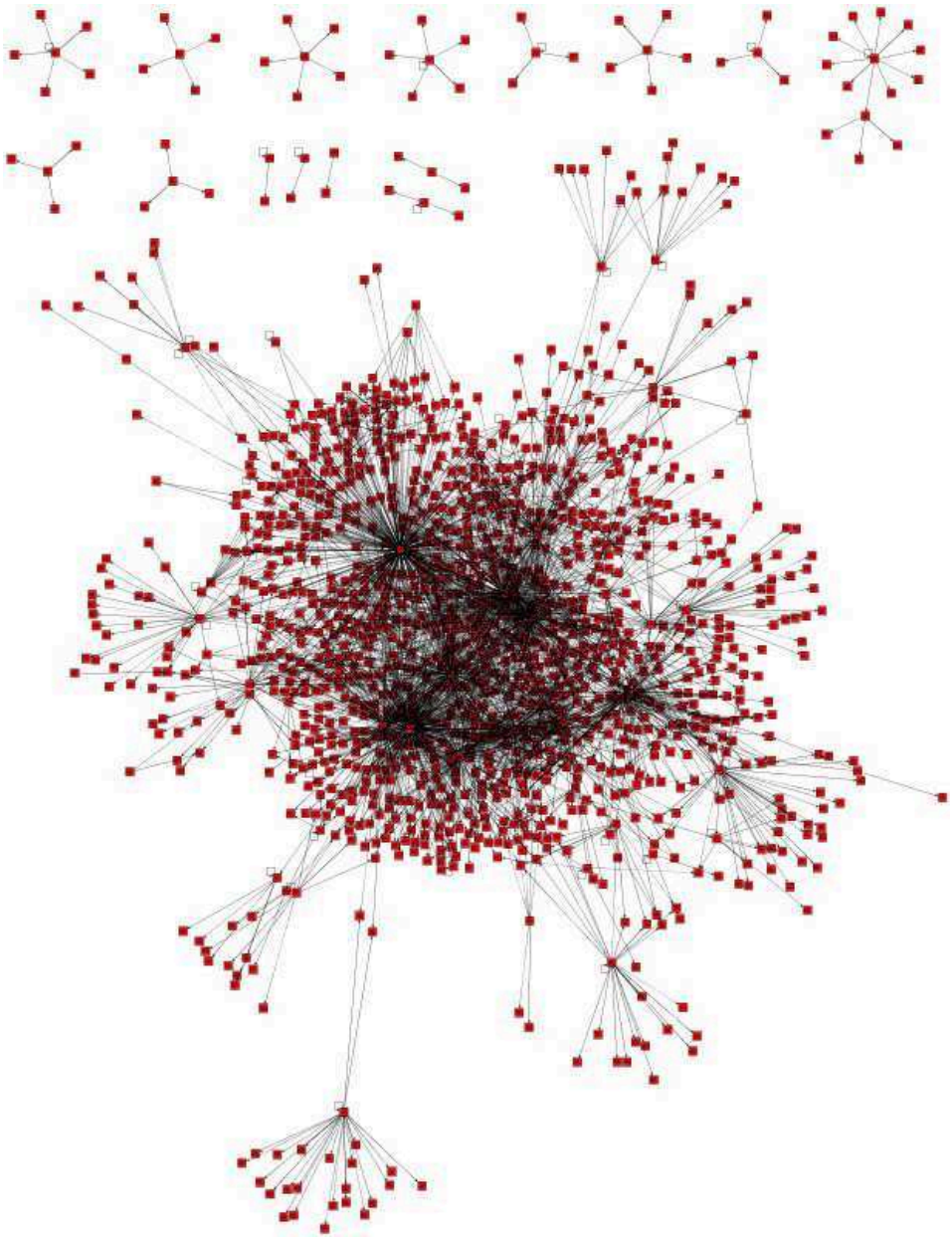

Figura B.2: Representación gráfica de la red de regulación del microorganismo *Escherichia coli*. Actualmente la red cuenta con 1,328 genes (RegulonDB ver. 5.5) conectados a través de 2,822 interacciones reguladoras, todas validadas experimentalmente [13] lo que la convierte en la red más completa que se conoce de un organismo vivo. La red se encuentra disponible en `http://regulondb.ccg.unam.mx`.

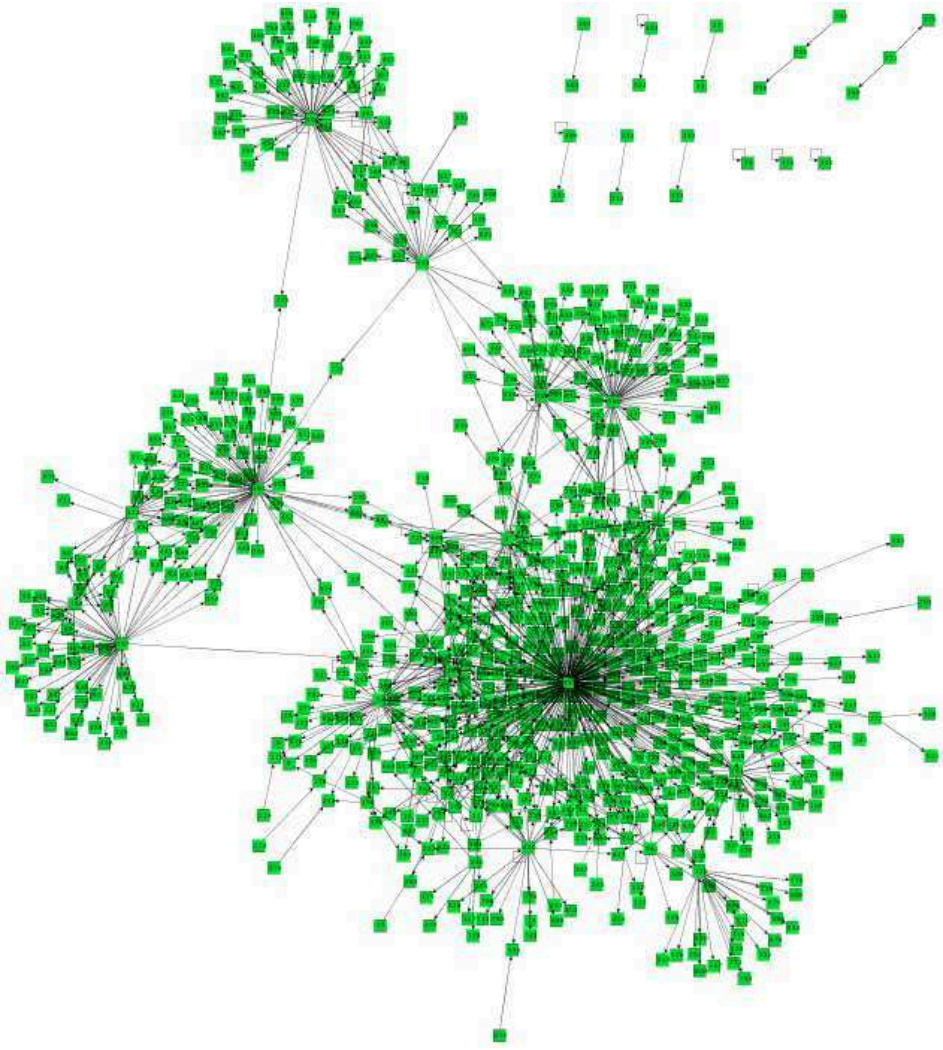

Figura B.3: Representación gráfica de la red de regulación del microorganismo *Bacillus subtilis*. La red cuenta con 830 genes y 1,267 interacciones reguladoras, todas ellas validadas experimentalmente [14]. La red se encuentra disponible en `http://dbtbs.hgc.jp`.

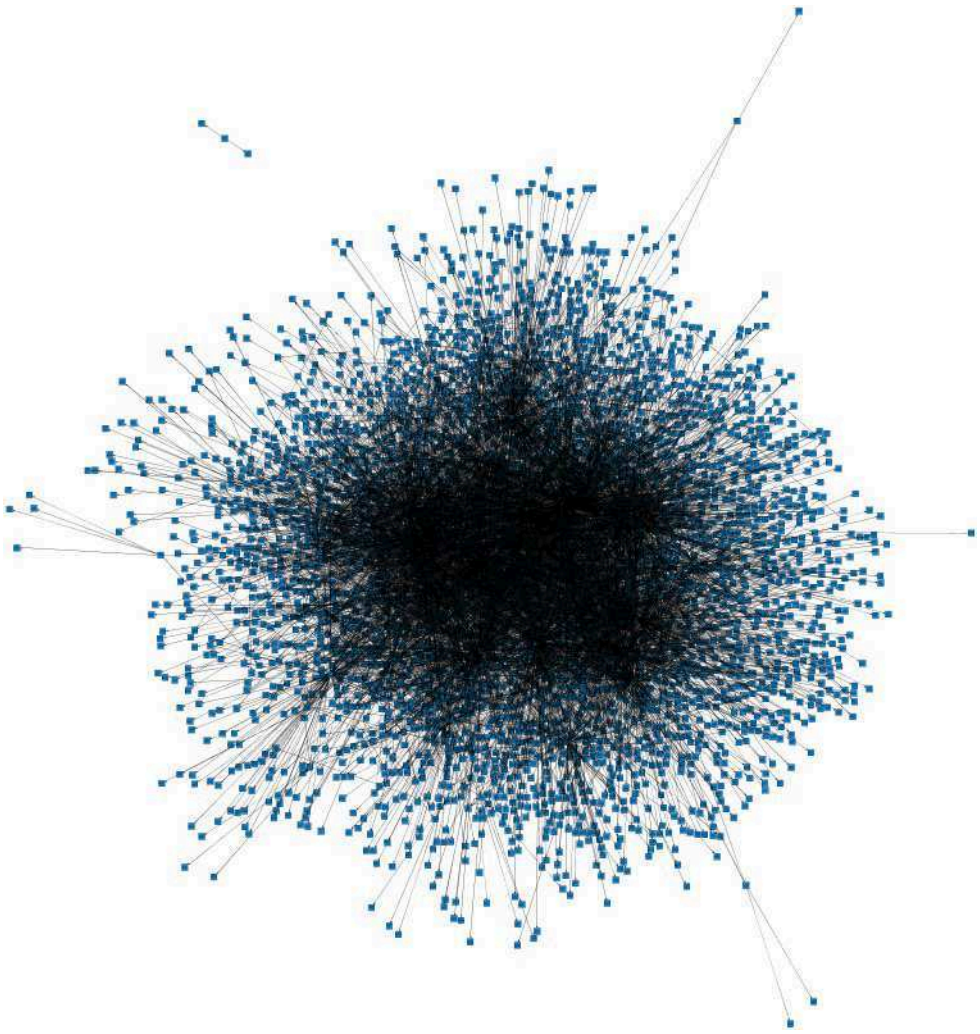

Figura B.4: Representación gráfica de la red de regulación del microorganismo *Saccharomyces cerevisiae*. La red cuenta con 3,459 genes conectados a través de 7,074 interacciones reguladoras. Parte de las interacciones en la red tienen validación experimental [12] el resto fue inferido usando experimentos ChIp-chip [99]. La unión de la red experimental e inferida fue hecha por Luscombe *et al.* [11] y se encuentra disponible en `http://sandy.topnet.gernsteinlab.org`.

regulados por ningún gen. Por la Fig. B.5 vemos que si no fuera por las asas de retroalimentación y los pocos cíclos de regulación, la estructura de control sería completamente jerárquica. Algo similar sucede con las redes de *B. subtilis* y *S. cerevisiae*.

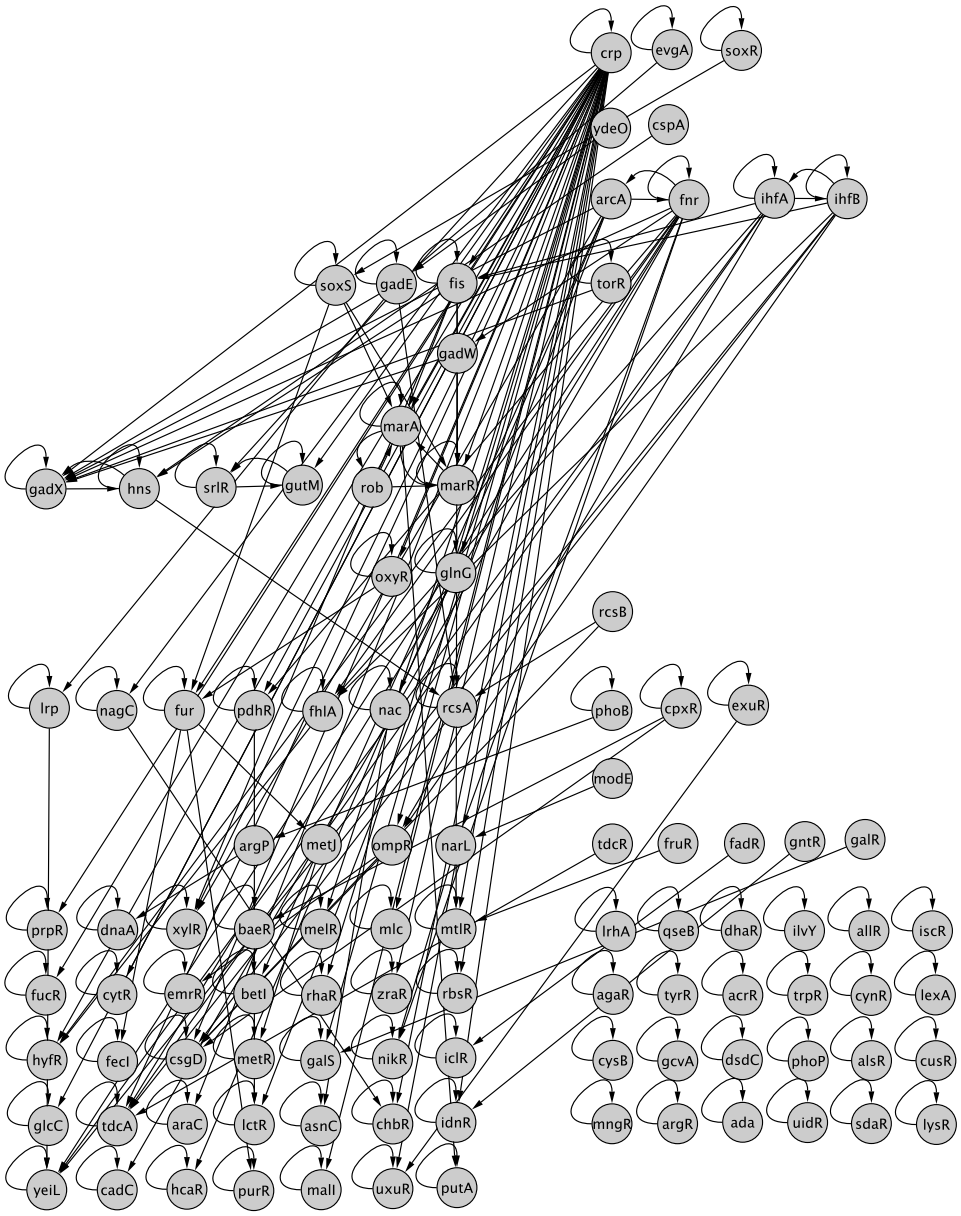

Figura B.5: Ordenamiento casi jerárquico de la cabeza de la red de regulación de *E. coli*. Genes en las capas inferiores pueden ser regulados por genes en las capas superiores o por genes *de su misma capa*. Genes en la mitad izquierda de las capas son regulados por al menos un gen de una capa superior. Genes en la mitad derecha no son regulados.

Bibliografía

[1] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403, 2000.

[2] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[3] C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler. Combinatorial synthesis of genetic networks. *Science*, 296:1466–1470, 2002.

[4] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827, 2000.

[5] J. Quackenbush01. Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427, 2001.

[6] S. Venkatasubbarao. Microarrays - status and prospects. *TRENDS in Biotechnology*, 22(12):630–637, 2004.

[7] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N.Ñannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. R. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, 2000.

[8] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M.Ñoble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, pages 840–846, 2006.

[9] H. Jeong, B. Tombor, R. Albert, Z.Ñ. Oltvai, and A.-L. Barabasi. The large scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[10] E. Almaas, Z.Ñ. Oltvai, and A.-L. Barabási. The activity reaction core and plasticity of metabolic networks. *PLoS Computational Biology*, 1(7):0557–0563, 2005.

[11] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[12] N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 2002.

[13] H. Salgado, S. Gamma-Castro, A. Martínez-Antonio, E. Díaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. García-Alonso, V. Jiménez-Jacinto, A. Santos-Zavaleta, C. Bonavidez-Martínez, and J. Collado-Vides. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization and growth conditions. *Nucleic Acids Research*, 34:(Database Issue:D394–7), 2006.

[14] Y. Makita, M.Ñakao, N. Ogasawara, and K.Ñakai. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Research*, 32:D75–77, 2004.

[15] G. R. Sambrano, G. Chandy, S. Choi, D. Decamp, R. Hsueh, K. M. Lin, D. Mock, N. O'Rourke, T. Roach, H. Shu, B. Sinkovits, M. Verghese, and H. Bourne. Unravelling the signal-transduction network in B-lymphicytes. *Nature*, 420:708–710, 2002.

[16] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 2005.

[17] J. M. G. Vilar, C. C. Guet, and S. Leibler. Modeling network dynamics: the *lac* operon, a case study. *The Journal of Cell Biology*, 161(3):471–476, 2003.

[18] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.

[19] S. Kauffman. *At Home in the Universe*. Oxford, 1995.

[20] C. Furusawa and K. Kaneko. Robust development as a consequence of generated positional information. *Journal of Theoretical Biology*, 224:413–435, 2003.

[21] R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118:4947–4957, 2005.

[22] A. L. Barabási. *Linked*. Plume, 2003.

[23] R. Lewin. Order for free. *New Scientist*, pages 10–11, 1993.

[24] S. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. of Theoret. Biol.*, 22:437–467, 69.

[25] R. Albert and H. G. Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.*, 223:1–18, 2002.

[26] C. Espinosa-Soto, P. Padilla-Longoria, and E. R. Alvarez-Buylla. A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *The Plant Cell*, 16:2923–2939, 2004.

[27] A. Chaos, M. Aldana, C. Espinosa-Soto, B. García Ponce de León, A. Garay Arroyo, and E. R. Álvarez-Buylla. From genes to flower patterns and evolution:dynamic models of gene regulatory networks. *Journal of Plant Growth Regulation*, 25:278–289, 2006.

[28] F. Jacob and J. Monod. Genetic repression, allosteric inhibition and cellular differentiation. In M. Locke, editor, *Cytodifferentiation and Macromolecular Synthesis*. Academic Press, New York, 1963.

[29] A. Wuensche and M. J. Lesser. *The Global Dynamics of Cellular Automata. An Atlas of Basin of Attraction Fields of One-Dimensional Cellular Automata*. Addison-Wesley, MA, 1992.

[30] A. Wuensche. *Modularity in Development and Evolution*, chapter 13, page 288. Chicago University Press, Chicago, 2004.

[31] J. M. W. Slack. Conrad Hal Waddington: The last Renaissance biologist? *Nature Reviews Genetics*, 3:889–895, 2002.

[32] S. Huang and D. Ingber. Shape-dependent control of cell growth, differentiation, and apoptosis: Switching between attractors in cell regulatory networks. *Experimental Cell Research*, pages 91–103, 2000.

[33] S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters*, page 128701, 2005.

[34] H. H. Chang, P. Y. Oh, D. E. Ingber, and S. Huang. Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell Biology*, 7(ll), 2006.

[35] R. Albert and H. G. Othmer. ...but no kinetic details are needed. *Siam News*, Dec. 2003.

[36] P. C. Hohenberg and B. I. Halperin. Theory of dynamical critical phenomena. *Rev. Mod. Phys.*, 49:435–479, 1977.

[37] D. L. Turcotte. Self-organized criticality. *Rep. Prog. Phys.*, 62:1377–1429, 1999.

[38] O. Peters and J. D. Neelin. Critical phenomena in atmospheric precipitation. *Nature Physics*, 2:393–396, 2006.

[39] A. Arakawa. Scaling tropical rain. *Nature Physics*, 2:373–374, 2006.

[40] T. Lux and M. Marchesi. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397:498–500, 1999.

[41] R.Ñ. Mantegna and H. E. Stanley. Scaling behavior in the dynamics of an economic index. *Nature*, 376:46–49, 1995.

[42] M. Heimpel. Critical behavior and the evolution of fault strenght during earthquake cycles. *Nature*, 388:865–868, 1997.

[43] C. H. Scholz. Earthquakes and friction laws. *Nature*, 391:37–42, 1998.

[44] S. Ostojic, E. Somfai, and B.Ñienhuis. Scale invarience and universality of force networks in static granular matter. *Nature*, 439:828–830, 2006.

[45] D. R. Chialvo. Are our senses critical. *Nature Physics*, 2:301–302, 2006.

[46] O. Kinouchi and M. Copelli. Optimal dynamical range of excitable networks at criticality. *Nature Physics*, 2006.

[47] G. Werner. Metastability, criticality and phase transitions in brain and its models. *BioSystems*, 2007.

[48] J. P. Sethna. Crackling noise. *Nature*, 410:242–250, 2001.

[49] P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks - methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62:247–292, 2000.

[50] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: *In Numero* molecular biology. *Nature Rev. Genet*, 2:268–279, 2001.

[51] M. E. Wall, W. S. Hlavacek, and M. A. Savageau. Design of gene circuits: Lessons from bacteria. *Nature Reviews Genetics*, 5:34–41, 2004.

[52] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[53] S. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.

[54] B. Derrida and D. Stauffer. Phase transitions in two dimensional Kauffman cellular automata. *Europhys. Lett.*, 2:739–745, 1986.

[55] M. Aldana. Boolean dynamics of networks with scale-free topology. *Physica D*, 185:45–66, 2003.

[56] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 1st edition edition, 2003.

[57] D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Corporation, 1996.

[58] J. Pearl. *Causality*. Cambridge University Press, 2000.

[59] J. W. Lengeler, G. Drews, and H. G.Schlegel, editors. *Biology of the Prokaryotes*. Blackwell Science, 1999.

[60] R. A. Irizarry and *et al.* Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):1–5, 2005.

[61] G. Rustici, J. Mata, K. Kivinen, P. Lió, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P.Ñurse, and J. Bähler. Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, 36:809–817, 2004.

[62] S. Kauffman. A proposal for using the ensemble aproach to understand genetic regulatory networks. *Journal of Theoretical Biology*, 230(4):581–590, 2004.

[63] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(824-827), 2002.

[64] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31:64–68, 2002.

[65] S. Mangan and U. Alon. Structure and function of the feed forward loop network motif. *Proc. Natl. Acad. Sci. USA*, 100:11980–11985, 2003.

[66] U. Alon. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8:450–461, 2007.

[67] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M.Ñitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, G. Sherlock, and C. A. Ball. The stanford microarray database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res*, 35(Database Issue):D766–770, 2007.

[68] G. Gosset, Z. Zhang, S.Ñayyar, W. A. Cuevas, and Jr. M. H. Saier. Transcriptome analysis of crp-dependent catabolite control of gene expression in *Escherichia coli*. *Journal of Bacteriology*, 186(11):3516–3524, 2004.

[69] Z. Zhang, G. Gosset, R. Barabote, C. G. Gonzalez, W. A. Cuevas, and Jr. M. H. Saier. Functional interactions between the carbon and iron utilization regulators, crp and fur, in *Escherichia coli*. *Journal of Bacteriology*, 187(3):980–990, 2005.

[70] R. Serra, M. Villani, and A. Semeria. Genetic network models and statistical properties of gene expression data in knock-out experiments. *Journal of Theoretical Biology*, pages 149–157, 2004.

[71] R. Serra, M. Villani, A. Graudenzi, and S. Kauffman. Why a simple model of genetic regulatory networks describes the distribution of avalanches in gene expression data. *Journal of Theoretical Biology*, 246:449–460, 2007.

[72] I. Shmulevich, S. A. Kauffman, and M. Aldana. Eukariotic cells are dynamically ordered or critical but not chaotic. *Proc. Natl. Acad. Sci. USA*, 102(38):13439–13444, 2005.

[73] T. Dobzhansky, F. J. Ayala, G. L. Stebbins, and J. W. Valentine. *Evolución*. Ediciones Omega, cuarta edition, 2003.

[74] S. Ohno. *Evolution by Gene Duplication*. Springer, New York, 1995.

[75] J. S. Taylor and J. Raes. Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.*, 38:615–43, 2004.

[76] J. Zhang. Evolution by gene duplication:an update. *TRENDS in Ecology and Evolution*, 18(6), 2003.

[77] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(1151-1155), 2000.

[78] M. Lynch. Gene duplication and evolution. *Science*, 297:945–947, 2002.

[79] M. Lynch and V. Katju. The altered evolutionary trajectories of gene duplicates. *TRENDS in Genetics*, 20(11):544–549, 2004.

[80] L. Margulis. *El Origen de la Célula*. Reverté, México, 2006.

[81] S. Huang. Back to the biology in systems biology: What can we learn from biomolecular networks? *Briefings in Functional genomics and proteomics*, 2(4):279–297, 2004.

[82] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of The Cell*. Garland Science, fourth edition edition, 2002.

[83] J. Stelling, U. Sauer, Z. Szallasi, and J. Doyle F. J. Doyle III. Robustness of cellular functions. *Cell*, 118:675–685, 2004.

[84] J. de Viser, J. Hermisson, G. P. Wagner, L. A. Meyers, H. Bagheri-Chaichian, J. L. Blanchard, and L. Chao. Perspective: Evolution and detection of genetic robustness. *Evolution*, 57(9):1959–1972, 2003.

[85] M. Kirschner and J. Gerhart. Evolvability. *PNAS*, 95:8420–8427, 1998.

[86] A. M. Poole, M. J. Phillips, and D. Penny. Prokaryote and eukaryote evolvability. *BioSystems*, 69:163–185, 2003.

[87] A. Wagner. *Robustness and Evolvability in Living Systems*. Princeton University Press, 2007.

[88] H. Kitano. Biological robustness. *Nature Reviews Genetics*, 5:826–837, 2004.

[89] A. Wagner. Robustness, evolvability and neutrality. *FEBS Lett.*, 2005b.

[90] S. Li, S. M. Assmann, and R. Albert. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, 4(10):1732–48, 2006.

[91] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichman. Structure and evolution of transcriptional regulatory networks. *Current Opinion In Structural Biology*, 14, 2004.

[92] I. Lozada-Chávez, S. C. Janga, and J. Collado-Vides. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Research*, 34(12):3434–3445, 2006.

[93] M. Aldana, S. Coppersmith, and L. P. Kadanoff. Boolean dynamics with random couplings. In E. Kaplan, J. E. Marsden, and K. R. Sreenivasan, editors, *Perspectives and Problems in Nonlinear Science*, pages 23–89. Springer, N. Y., 2003.

[94] I. Yanai, C. J. Camacho, and C. DeLisi. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Physical Review Letters*, 85(12):2641–2644, 2000.

[95] A. M. Evangelisti and A. Wagner. Molecular evolution in the yeast transcriptional regulation network. *J. Exp. Zool.*, 302B:392–411, 2004.

[96] S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36:492–496, 2004.

[97] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406:188–192, 2000.

[98] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 2003.

[99] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[100] E. Segal, M. Shapira, A. Regev, D. Peér, D. Botstein, D. Koller, and N. Friedman. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

[101] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 2005.