MEMORIAS DE LA XIX ESCUELA DE VERANO DE FÍSICA

julio 25-agosto 5, 2011

Editores

José Récamier
Rocío Jáuregui
Manuel Torres

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
2012

Esta edición fue preparada por el Instituto de Física
y el Instituto de Ciencias Físicas de la UNAM.

Primera edición electrónica: 2012

ISBN:

Hecho en México

CONTENIDO

INTRODUCCIÓN

La XIX Escuela de Verano en Física fue organizada por el Posgrado en Ciencias Físicas, el Instituto de Física y el Instituto de Ciencias Físicas de la Universidad Nacional Autónoma de México. Se llevó a cabo en las instalaciones del Instituto de Física, en Ciudad Universitaria, del 25 al 29 de julio de 2011 y en las instalaciones del Instituto de Ciencias Físicas, en Cuernavaca, Morelos, del 1 al 5 de agosto de ese año.

En esta escuela se impartieron ocho cursos de cinco horas de duración cada uno y 25 conferencias. Los cursos y conferencias cubrieron un amplio espectro con temas como cosmología, plasmas, biofísica, óptica de metamateriales, nanobiotecnología, caminatas aleatorias, mecánica cuántica, cuerdas, sistemas complejos y biología teórica entre otros.

Rocío Jáuregui, Instituto de Física
José Récamier, Instituto de Ciencias Físicas
Manuel Torres, Instituto de Física
Rafael Pérez Pascual, Instituto de Física
Universidad Nacional Autónoma de México
Ciudad Universitaria, marzo, 2012

AGRADECIMIENTOS

Agradecemos a Yanalté Herrero por el apoyo secretarial y a Ulises Amaya por el apoyo con la página de la escuela.

Agradecemos los apoyos recibidos para la realización de esta escuela a la Universidad Nacional Autónoma de México a través de la Coordinación de la Investigación Científica, del Instituto de Física (IF) y del Instituto de Ciencias Físicas (ICF). Agradecemos también al programa de Becas de Movilidad Santander-Universia y a las Becas ECOES.

PROFESORES PARTICIPANTES

- Maximino Aldana González, ICF, *Qué son los sistemas complejos y cómo se comportan*.
- Miguel Ángel Ávila, Unidad PET/Ciclotrón, Facultad de Medicina, *Tomografía por emisión de positrones*.
- Pablo Barberis, IIMAS, *Manipulación de batimientos cuánticos en Electrodinámica Cuántica de cavidades*.
- María Ester Brandán, IF, *Resta de imágenes en mamografía*.
- Remigio Cabrera Trujillo, ICF, *Resonancias de Feshbach*.
- Bernardo Campillo Illanes, FQ, *Plasmas*.
- Juan Carlos Cheang, IF, *Introducción a las nanociencias: materiales, métodos de síntesis y de caracterización*.
- Jaime de Urquijo Carmona, ICF, *Plasmas de baja temperatura: Esencia y aplicaciones*.
- Rubén Fossion, Instituto de Geriatría, Secretaría de Salud, *Fractales y caos en la medicina: la hipótesis de pérdida de complejidad con enfermedades y envejecimiento*.
- Gerardo García Naumis, IF, *Problemas abiertos del estado sólido y la física de materiales*.
- Gabriel Germán Velarde, ICF, *Cosmología del universo temprano*.
- Shahen Hacyan, IF, *Mecánica cuántica: fundamentos y aplicaciones*.
- Antonio Juárez Reyes, ICF, *Métodos avanzados de cavidades ópticas en física molecular y sus aplicaciones*.
- Hernán Larralde Ridaura, ICF, *Caminatas aleatorias: Desde física hasta finanzas*.
- Eugenio Ley Koo, IF, *Momento angular en bases de armónicos esféricos y esferoconales*.
- Horacio Martínez Valencia, ICF, *Plasmas*.
- Gustavo Martínez Mekler, ICF, *Redes lógicas relacionadas con el nado de espermatozoides de erizo de mar*.
- Guerda Masillon, IF, *La física: pieza clave en el diagnóstico y tratamiento de enfermedades*.
- Luis Alberto Medina, IF, *Dosimetría interna en la terapia con radionúclidos*.
- W. Luis Mochán Backal, ICF, *Respuesta óptica de metamateriales*.
- Iván Ortega Blake, ICF, *Fisicoquímica de transporte transmembranal*.
- Nina Pastor Colón, FC-UAEM, *Viaje por el paisaje conformacional de una proteína con doble personalidad*.
- Rafael Pérez Pascual, IF, *Caos y física no lineal*.

- Carlos Pineda, IF, *Información cuántica*.
- Pedro Quinto Su, ICN, *Aplicaciones de holografía digital a manipulación óptica*.
- Fernando Ramírez, ICN, *Introducción a las técnicas de enfriado, atrapamiento y manipulación de átomos neutros*.
- Saúl Ramos, IF, *Fenomenología de cuerdas*.
- José Récamier Angelini, ICF, *Estados coherentes generalizados*.
- Alejandro Reyes, IF, *Plasmónica: una ruta hacia la manipulación de la luz en una escala nanométrica*.
- Luis Rodríguez, IF, *Efectos de la irradiación de sólidos con iones*.
- José Luis Ruvalcaba, IF, *Técnicas espectroscópicas para el estudio no destructivo del patrimonio cultural*.
- Humberto Saint Martin, ICF, *Taller de biofísica molecular computacional*.
- Fco. Javier Sevilla, IF, *La relación fluctuación-disipación de Física Estadística*.
- Julia Tagüeña Parga, CIE, *De la electrónica a la fotónica*.
- Alfred U'Ren, ICN, *Generación de luz no clásica con propiedades diseñadas a la medida*.
- Gabriel Vázquez Torres, ICF, *Espectroscopia electrónica del HCl*.
- Carlos Villarreal Luján, IF, *Redes genéticas en Biología y Medicina*.
- Karen Volke, IF, *Interacción de micropartículas con potenciales ópticos periódicos*.

ALUMNOS PARTICIPANTES

- Juan Carlos Alejandre Vázquez
- Anahí Alvarado Sánchez
- Zurika Iveth Blanco García
- Carlos Bracamontes Palma
- Durán Mississippi Valenzuela
- Nephtalí Garrido González
- Lucía Gómez Córdova
- Santiago González Larios
- Jorge Guzmán Maldonado
- Ramón Heberto Martínez Mayorquin
- Manuel Mijaíl Martínez Ramos
- César Martínez Robles
- Esteban Martínez Vargas
- Manuel Mendoza López
- Lili Jazmín Pereyra Carballal
- Diana Consuelo Pérez Pérez
- Reinher Rolando Pimentel Domínguez
- Irving Enrique Reyna Nolasco
- Cyndi Diana Rodríguez García
- Edgar Omar Rodríguez Rojas
- Alejandro Rosado Fuentes
- Víctor Sánchez Cordero Canela
- Leticia Somera León
- Sergio Camilo Vargas Ávila
- Diana Carolina Vargas Ortega
- Ana Karem Vega Salgado
- Luis Armando Vieyra Rebollo

Redes Complejas

Maximino Aldana

Diciembre 2011

"La materia no sólo interactúa, también se organiza. Conocemos básicamente todas las leyes de interacción de la materia, pero no sabemos casi nada sobre sus leyes de organización".

Albert L. Lehninger (1917-1986)

1. Introducción

En los últimos 8 años hemos sido testigos de una explosión en el estudio de las propiedades estructurales y dinámicas de las redes complejas. Durante este tiempo se han publicado cientos de artículos sobre este tema en revistas de investigación científica internacionales de diferentes disciplinas, que abarcan física, biología, sociología, neurología, economía, medicina, por mencionar algunos ejemplos. Este interés en las redes complejas radica en que nos hemos dado cuenta de que dichas redes abundan en la naturaleza, son parte de nuestra vida diaria y se presentan a diferentes niveles de organización. Por ejemplo, algunas redes biológicas que encontramos en el nivel microscópico son las *redes de regulación genética*, *redes de proteínas*, *redes neuronales*, *redes metabólicas*. Por otro lado, a un nivel de organización mucho mayor, encontramos *redes de comunicación e informáticas* (la red internet, la red www, redes telefónicas, etc.), *redes sociales* (amistades, contactos sexuales, colaboradores científicos, propagación de enfermedades, etc.), *redes ecológicas* (interacciones tróficas en un ecosistema). Las redes complejas son ubicuas, están por todos lados. Incluso se ha estudiado la red de super héroes en el Universo de Marvel, siendo el hombre ara na el super héroe más popular con la mayor conectividad [1]. Es un hecho sobresaliente el que todas estas redes, tan diferentes en naturaleza y en tama no, tengan muchas propiedades

estructurales similares. Este hecho, tan simple como sorprendente, hace posible que podamos formular modelos matemáticos para entender y explicar las propiedades estructurales (y en algunos casos también las propiedades dinámicas) de las redes complejas.

En estas notas presento algunas de las herramientas matemáticas y conceptuales que se han desarrollado a lo largo de varios a nos para analizar la estructura y dinámica de las redes complejas. Los que estén interesados podrán encontrar un estudio mucho más completo del que presento aquí en las referencias [2, 3, 4, 5], las cuales son trabajos de revisión excelentes que sirven como punto de partida para adentrarse en el fascinante mundo de las redes complejas.

2. Algunas definiciones

2.1. Sistemas complejos

Comenzemos nuestro estudio describiendo qué son los sistemas complejos. Como ocurre con la gran mayoría de los conceptos científicos, no podemos definir los sistemas complejos en un simple enunciado. En lugar de ésto, vamos a enumerar las características más importantes que son comunes a todos los sistemas complejos:

1. Están compuestos de muchas partes que interactúan entre sí. De hecho, el adjetivo "Complejo" en este contexto no significa solamente que el sistema sea complicado, sino también que está compuesto de muchas partes, como un *complejo* industrial.

2. Cada parte tiene su propia estructura interna y está encargada de llevar a cabo una función específica.

3. Lo que ocurra a una parte del sistema afecta de manera **altamente no lineal** a todo el sistema.

4. Presentan **comportamientos emergentes**, de tal manera que el *el todo no es la simple suma de sus partes.*

Como un ejemplo típico de sistema complejo consideremos a la célula. Evidentemente la célula está compuesta de muchas partes (ribosomas, mitocondrias, núcleo, membrana, retículo endoplasmático, ADN, ARN, etc.),

y cada una de estas partes se encarga de realizar alguna función específica dentro de la célula. Las partes de la célula responden de forma no lineal ante perturbaciones externas. Por ejemplo, algunas veces una mutación en el ADN no tiene ningún efecto en la célula, mientras que otras veces una sóla mutación puede ser fatal[1]. Además, la célula presenta comportamientos emergentes que no pueden explicarse en términos de las propiedades de sus partes individuales. Así, podemos hablar de una célula enferma, pero no podemos decir que un ribosoma o una proteína estén enfermos. La enfermedad es una propiedad que emerge como resultado de la organización colectiva de todos los constituyentes de la célula.

2.2. Redes Complejas

Las redes complejas son conjuntos de muchos nodos conectados que interactúan de alguna forma. A los nodos de una red también se les llama *vértices* o *elementos* y los representaremos por los síimbolos v_1, v_2, ..., v_N, donde N es el número total de nodos en la red. Si un nodo v_i está conectado con otro nodo v_j, esta conexión se representa por una pareja ordenada (v_i, v_j). La definición matemática de una red (también llamada *grafo* por los matemáticos) es la siguiente:

Definición 1 *Una red $\mathcal{R}$ consiste de un conjunto de nodos $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$, y un conjunto de parejas ordenadas $\mathcal{E} = \{(v_i, v_j)\} \subset \mathcal{V} \times \mathcal{V}$. Cada pareja ordenada (v_i, v_j) se llama* **conexión dirigida** *del nodo v_i al nodo v_j. La red $\mathcal{R}$ se llama* **no dirigida** *si para cada pareja $(v_i, v_j) \in \mathcal{E}$ también existe la pareja $(v_j, v_i) \in \mathcal{E}$. De lo contrario, la red se denomina* **dirigida**. *Llamaremos a todos los nodos que estén conectados directamente a un nodo v_i, los* **vecinos** *de v_i. Finalmente, el número k_i de vecinos del nodo v_i (es decir, el número de conexiones de v_i) se llama la* **conectividad** *de v_i, y el promedio de estas conectividades, $\langle k \rangle = N^{-1} \sum_{i=1}^{N} k_i$, es la* **conectividad de la red**.

Aunque la definición formal de una red es útil en el desarrollo matemático de la teoría, para nuestros propósitos basta con considerar que una red es un montón de nodos entre los que existen coneciones. En la naturaleza se

[1]La *anemia falciforme*, por ejemplo, es una enfermedad de los glóbulos rojos de la sangre que se origina por una sóla mutación en uno de los más de 600 aminoácidos que conforman a la proteína β-globina encargada de capturar oxígeno.

Redes sociales	
Sexuales	Dos personas están conectadas si han tenido por lo menos una relación sexual
Actores	Dos actores están conectados si han aparecido en la misma película
Amistades	Dos personas están conectadas si son amigas
Científicos	Dos científicos están conectados si han sido coautores en algún artículo
Familiares	Dos personas están conectadas si son familiares cercanos
Enfermedades	Dos personas están conectadas si una contagió de una enfermedad a la otra
Redes informáticas	
Internet	Dos computadoras están conectadas si hay un cable que las conecta
WWW	Dos páginas web están conectadas si hay un hipervínculo de una a la otra
Palabras	Dos palabras están conectadas si en el diccionario una aparece en la definición de la otra
Palabras	Dos palabras están conectadas si son sinónimos

Cuadro 1: Diferentes tipos de redes.

pueden encontrar muchos tipos de redes, es decir, muchos tipos de nodos y conexiones. Por ejemplo, en una red social los nodos son las personas y las conexiones pueden ser los lazos de amistad que existan entre ellas: dos personas están conectadas si son amigos. En la misma sociedad podemos definir las conexiones de forma distinta, por ejemplo, dos personas están conectadas si han tenido relaciones sexuales. Claramente, la red definida a través de amistades es diferente a la red definida a través de contactos sexuales, ya que el hecho de que dos personas sean amigas no significa que hayan tenido relaciones sexuales, y viceversa.

Notemos entonces que incluso en un mismo conjunto de nodos podemos definir redes diferentes dependiendo de como hayamos definido las conexiones, lo cual, por supuesto, depende del fenómeno que nos interese estudiar.

Redes biológicas	
Protéicas	Dos proteínas están conectadas si participan en la misma reacción química
Genéticas	Dos genes están conectados si uno regula la expresión del otro
Ecológicas	Dos especies están conectadas si una se come a la otra
Neuronales	Dos neuronas están conectadas si existe una conexión sináptica entre ellas

Cuadro 2: Diferentes tipos de redes (continuación del cuadro anterior).

Por ejemplo, si estuviésemos interesados en analizar como se propaga una enfermedad como el SIDA en una sociedad, claramente nos convendría estudiar la red de interacciones sexuales, mientras que si estamos interesados en encontrar a un asesino, lo que nos conviene es estudiar la red de amistades, ya que son sus amigos los que pueden darnos información sobre su paradero. Los cuadros 1 y 2 muestran diferentes tipos de redes que se encentran en la naturaleza.

En los ejemplos anteriores hay redes dirigidas y redes no dirigidas. Por ejemplo, la red de contactos sexuales es *no dirigida*, ya que si A tuvo relaciones con B, entonces evidentemente B tuvo relaciones con A. Pero la red de transmisión de la gripe es *dirigida*, ya que si A contagió de gripe a B, no necesariamente B contagió también a A. Otra red dirigida es la World Wide Web (WWW). En mi página web yo tengo una liga a la página del periódico *la jornada*, pero en la página de *la jornada* no hay ninguna liga a mi página web. En este sentido, hay una conexión de mi página hacia la de *la jornada*, pero no hay una conexión de regreso.

Intuitivamente, una red no dirigida puede pensarse como aquella en la que las conexiones entre los nodos siempre son simétricas (si A está conectado con B, entonces B está conectado con A), mientras que en una red dirigida no todas las conexiones son simétricas, es decir, siempre existen conexiones asimétricas (A está conectado con B pero B no está conectado con A).

Otro concepto importante es el de *islas* (o *sub redes*) de una red. Notemos que la definición de red que dimos arriba **no dice** que todos los nodos deben estar conectados unos con otros. Ni siquiera dice que todos los nodos deben tener conexiones. La definición matemática nos dice que una red es

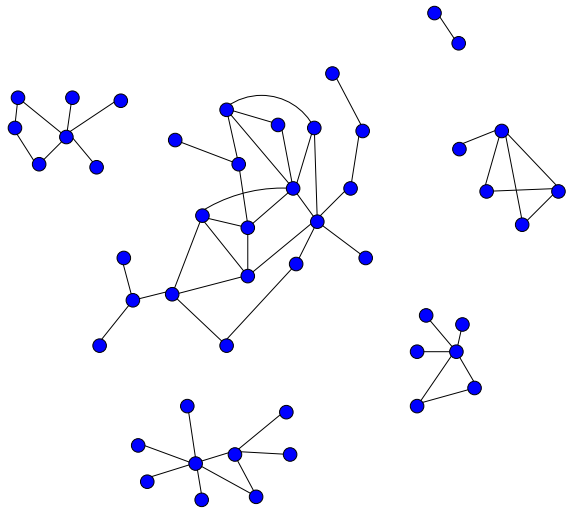


Figura 1: Una red puede estar compuesta de varias islas, como en el ejemplo mostrado en esta figura. La isla más granda se denomina la *isla gigante*.

un conjunto de nodos entre los que existen *algunas* conexiones. Esto quiere decir que en la red pueden existir nodos que no tengan conexiones, es decir, nodos aislados. También pueden existir grupos de nodos que estén conectados entre sí pero que no estén conectados con el resto de la red. Como un ejemplo concreto pensemos en una red social en la que dos individuos están conectados sin son familiares cercanos (específicamente, hermanos, medios hermanos, primos, padres, hijos, esposos, tíos, sobrinos, abuelos y nietos). Esta es evidentemente una red no dirigida, ya que si A es familiar de B, entonces B también es familiar de A. Sin embargo, muy probablemente en una sociedad grande esta red estaría fracturada en islas o sub redes, debido a que claramente no todas las personas en una sociedad son familiares cercanos de todos los demás. En mi caso particular, mi apellido es "Aldana González", por lo que es natural pensar que estoy conectado directamente con algunas de las familias Aldana y algunas las familias González en México. Pero muy probablemente ningún miembro de mi familia está conectado con alguien de la familia Azcárraga, o con la familia Zabludovsky. Por lo tanto, los miembros de mi familia conforman una isla o sub red dentro de la cual estamos conectados entre nosotros, pero esta isla está desconectada de otras familias de la sociedad. La Fig. 1 muestra un ejemplo de una red compuesta de varias islas.

Las islas en una red pueden tener diferentes tama nos, que van desde 1 (un sólo nodo que no está conectado a nadie) hasta el tama no de toda la red (todos los nodos están conectados con todos), en cuyo caso la red consiste de una sóla isla, que es ella misma. Es importante enfatizar que el hecho de que una isla no esté conectada al cuerpo principal de la red no significa que dicha isla no pertenezca a la red. La red no está determinada sólo por las conexiones, sino también por los nodos que conforman al sistema. Esto puede parecer poco intuitivo, pero desde el punto de vista matemático es conveniente considerar que todos los nodos del sistema pertenecen a la red, independientemente de que haya o no conexiones entre ellos. Como veremos más adelante, las islas juegan un papel importante en la teoría de redes.

3. Estudio de las redes complejas

Ejemplos de redes dirigidas y no dirigidas abundan en la naturaleza. Ahora ustedes pueden comenzar a pensar en la red que más les guste, ya sea una red biológica, social, informática o cualquier otra. Las redes se presentan en diferentes tamaos, colores y sabores. Pero, ¿qué hacemos con las redes? ¿Cómo las estudiamos?

El estudio general de las redes complejas puede dividirse en dos campos diferentes y complementarios: *Estructura* y *Dinámica* (ver la Fig. 2). En el primer campo de estudio uno está interesado en determinar las propiedades estructurales (o topológicas) de la red, es decir, en las propiedades que nos dicen cómo están conectados los nodos unos con otros. Algunas de las propiedades más importantes que determinan la estructura (o topología) de una red son las siguientes:

1. **La distribución de conexiones (o vecinos)** $P(k)$: Es la probabilidad de que un nodo escogido al azar tenga k conexiones (o vecinos). Por ejemplo, en una red de contactos sexuales $P(k)$ es la probabilidad de que una persona escogida al azar en una sociedad haya tenido k parejas sexuales distintas a lo largo de su vida.

2. **El coeficiente de agregación** C: Es la probabilidad de que dos nodos conectados directamente a un tercer nodo, estén conectados entre sí (ver la Fig. 3(a)). Por ejemplo, en una red de amistades, es la probabilidad de que dos de mis amigos sean ellos mismos amigos uno del otro.

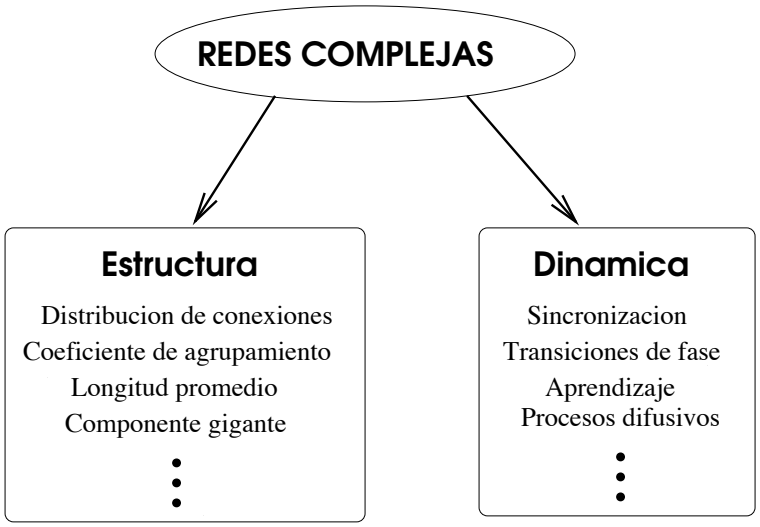


Figura 2: El estudio de las redes complejas puede dividirse en dos partes: (a) el estudio de sus propiedades estructurales y (b) el estudio de sus propiedades dinámicas.

3. **La longitud mínima L_{ij} entre dos nodos v_i y v_j**: Es el número mínimo de "brincos"que se tienen que dar para llegar de un nodo v_i de la red a otro nodo v_j de la red. Por ejemplo, en la red mostrada en la Fig. 3(b), aunque existen varios caminos para llegar de v_i a v_j, el camino mínimo consiste de tres pasos (indicado con líneas gruesas).

4. **La longitud promedio de la red L**: Es el promedio de las longitudes mínimas L_{ij} entre todas las posibles parejas de nodos (v_i, v_j) de la red.

5. **La distribución de tama nos de islas $P(s)$**: Es la probabilidad de que una isla esté compuesta por s nodos.

6. **El tama no de la isla más grande**, al que denotaremos por S_∞.

En una red, los nodos además de estar conectados también interactuan, y las interacciones pueden dar lugar a fenómenos dinámicos muy interesantes. Por lo tanto, además de estudiar las propiedades estructurales de una red también es importante estudiar sus propiedades dinámicas una vez que sabemos de qué manera interactuan los nodos. Por ejemplo, las enfermedades en una sociedad no son estáticas, sino que se propagan por toda la población

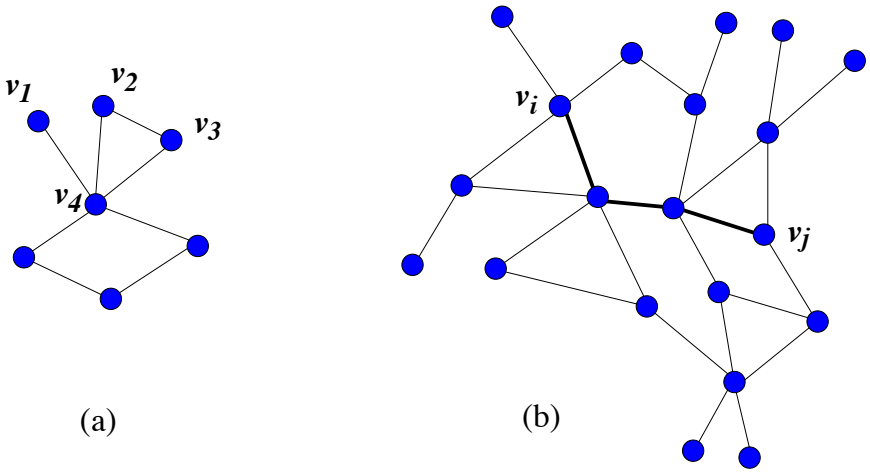


Figura 3: (a) Los nodos v_1, v_2 y v_3 están conectados al nodo v_4. Sin embargo, los nodos v_1 y v_2 no están conentados entre sí, mientras que los nodos v_2 y v_3 sí lo están. Esto significa que no todos los "amigos" de v_4 son amigos entre sí, lo cual disminuye el coeficiente de agregación. (b) Aún cuando existen varios caminos para llegar del nodo v_i al nodo v_j, el camino de longitud mínima consiste de tres pasos, indicados con líneas gruesas en la figura.

dando lugar a epidemias. Las neuronas en el cerebro están conectadas físicamente unas con otras por medio de las uniones entre dendritas y axones. A través de dichas uniones las neuronas se transmiten se nales eléctricas que se propagan por todo el cerebro y que dan lugar a una serie de fenómenos dinámicos interesantísimos, entre los cuales destacan el reconocimiento de imágenes y sonido, la motricidad de los músculos, el lenguage, el pensamiento y finalmente la consciencia. Otros ejemplos son la propagación de virus informáticos en la red internet, o la comunicación entre los peces que da lugar a grupos enormes de peces moviendose todos en la misma dirección. En fin, existen tantos fenómenos dinámicos en redes complejas como interacciones físicas, químicas, informáticas, o sociales se puedan imaginar, y cada día aparcen más y más artículos en la literatura científica donde se estudian nuevos procesos dinámicos sobre redes complejas.

En estas notas introductorias nos enfocaremos más al estudio de las propiedades estructurales de las redes complejas. Sin embargo, en el último capítulo veremos un poco las propiedades dinámicas de redes neuronales.

4. Estructura de las redes complejas

4.1. Distribución de vecinos

Tal vez la propiedad más importante que caracteriza la estructura de una red compleja es la distribución de vecinos $P(k)$, que nos dá la probabilidad de que un nodo escogido al azar tenga k conexiones (o vecinos). En los trabajos recientes que se han llevado a cabo para caracterizar a las redes complejas se ha encontrado que existen tres tipos de distribuciones $P(k)$ importantes, las que determinan tres estructuras o topologías[2] diferentes:

$$\text{Topología de Poisson} \quad P(k) = e^{-z}\frac{z^k}{k!}, \tag{1}$$

$$\text{Topología Exponencial} \quad P(k) = Ce^{-\alpha k}, \tag{2}$$

$$\text{Topología Libre de Escala} \quad P(k) = Ck^{-\gamma}. \tag{3}$$

Las redes con topología de Poisson son importantes principalmente por razones históricas, ya que dichas redes fueron las primeras que se analizaron matemáticamente. Este análisis lo llevaron a cabo los matemáticos húngaros Paul Erdös (1913-1996) y Alfréd Rényi (1921-1970) en la década de los 50s. Ellos también reportaron la primera transición de fase topológica observada en redes con topología de Poisson. Por lo tanto, a estas redes también se les conoce como redes tipo Erdös-Rényi. Sin embargo, a pesar de su importancia histórica, las redes con topología de Poisson están lejos de ser una representación realista de las redes reales observadas en la naturaleza. No fue sino hasta 1998 que se comenzó el estudio sistemático de las propiedades topológicas de las redes complejas reales. En este estudio participaron principalmente y de forma independiente Albert Lásló-Barabási, Ricard Solé y Mark J. Newman. Ellos encontraron que la topología exponencial aparece algunas veces en las redes reales. Pero el resultado más sorprendente de sus estudios fue la ubicuidad de las redes con topología libre de escala, la cual aparece prácticamente

[2]En el contexto de las redes complejas, la palabra "topología"no significa lo mismo que en análisis funcional. Aquí "topología" es sinónimo de "estructura" o "arquitectura".

Red	**Núm. de nodos**	**Núm. de conexiones**	γ_i
dominio www.nd.edu	325,729	1,469,680	2.1
Páginas de WWW encontradas por Altavista	$2{,}711 \times 10^9$	$2{,}130 \times 10^9$	2.1
Dominos en la WWW	$2{,}60 \times 10^5$	——	1.9
Nivel de inter-dominio de la Internet	4,389	8,256	2.2
Sistemas autónomos en la Internet	6,374	13,641	2.2
Nivel de ruteador en la Internet	150,000	200,000	2.3
Citas en la base de datos ISI	783,339	6,716,198	3.0
Citatas de la revista *Phys. Rev. D*	24,296	351,872	2.3
Red de colaboraciones de actores de Hollywood	212,250	61,085,555	2.3

Cuadro 3: Algunas de las redes libres de escala que se han encontrado en la naturaleza. Sólo se muestra el exponente de entrada para las redes dirigidas. Los cuadros con líneas "——ïndican que yo no tenía el dato correspondiente al momento de escribir estas notas.

en todos lados, desde las peque nas redes metabólicas dentro de la célula, hasta las grandes redes informáticas como la red Internet. En los cuadros 3 y 4 se listan algunas de las redes con topologías libres de escala que se han encontrado en los últimos 10 a nos.

Seguramente se estarán preguntando qué es lo sorprendente respecto a la topología libre de escala. Bueno, por un lado sorprende que esta topología se encuentre en redes tan diferentes y de tan gran variedad como las listadas en los cuadros 3 y 4. El hecho de que la topología libre de escala aparezca por todos lados sugiere que podría existir un mecanismo simple que genera este tipo de redes a diferentes niveles de organización, desde las peque nas redes intracelulares hasta las grandes redes sociales o informáticas. ¿Podría ser esto posible? ¿Será cierto que la formación de redes tan diferentes como la red de interacciones protéicas de *S. cerevisiae* (levadura), la red Internet y la red

Red	**Núm. de nodos**	**Núm. de conexiones**	γ_i
Red de colaboradores en las revistas *Medline*	1,388,989	$1{,}028 \times 10^7$	2.5
Colaboradores en las revistas de matemáticas	70,975	$1{,}32 \times 10^5$	2.1
Colaboradores an las revistas de neurociencias	209,293	$1{,}21 \times 10^6$	2.4
Red de interacciones metabólicas en (*E. coli*)	778	$\sim 1500 - 3000$	2.2
Red de interacciones protéicas en levadura	1,870	2,240	2.5
Co-ocurrencia de palabras	470,000	17,000,000	2.7
Red de palabras sinónimas	22,311	—	2.8
Circuitos digitales	2×10^4	4×10^4	3.0
Llamadas telefónicas	47×10^6	8×10^7	2.1
Red de interacciones sexuales en humanos	2,810	—	3.4
Redes alimenticias (interacciones tróficas)	154	405	1.0

Cuadro 4: Continuación del cuadro 3.

de colaboraciones científicas en las revistas de neurociencias, esté governada por la misma ley fundamental? No lo sabemos aún.

Por otro lado, la topología libre de escala es sorprendente porque no se esperaba que existiera. En la naturaleza existen muchos procesos aleatorios que generan distribuciones de Poisson o distribuciones exponenciales, pero existen muy pocos procesos conocidos que generan distribuciones libres de escala como la dada en la Eq. (3)[3]. De hecho, el trabajo de Erdös y Rényi demostró que las redes que se construyen a nadiendo nuevos nodos y conexiones al azar presentan topologías de Poisson o exponenciales. Y ésta es precisamente la paradoja, que la red internet, la red de colaboraciones científicas y la red de contactos sexuales, por ejemplo, son redes que se formaron aleatoriamente a nadiendo nuevos nodos y nuevas conexiones a lo largo del tiempo. Entonces, ¿cómo es posible que estas redes que se formaron al azar no presenten topologías de Poisson o topologías exponenciales como lo habían predicho

[3]A las distribuciones libres de escala también se les llama distribuciones de potencia.

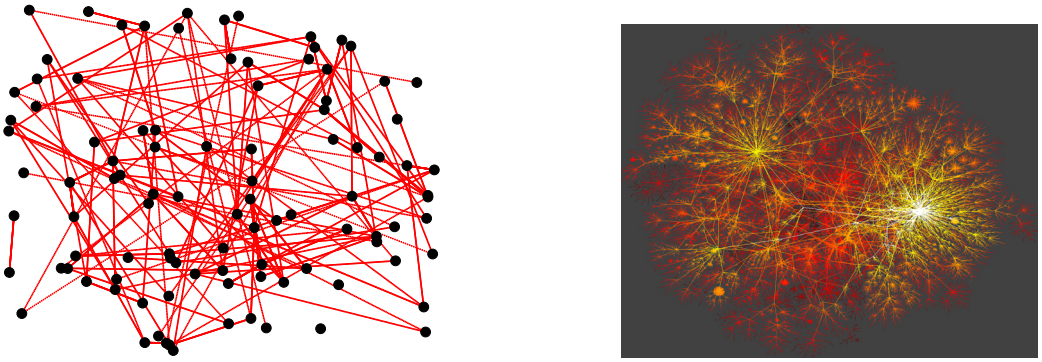


Figura 4: La figura de la izquierda muestra una red aleatoria con topología de Poisson. La red de la derecha es la red Internet al nivel de ruteadores, la cual es una red con topología libre de escala.

Erdös y Rényi?

Las redes con topología de Poisson son muy diferentes estructuralmente a las redes con topología libre de escala. La Fig. 4 muestra una red de Poisson (izquierda) y una red libre de escala (derecha). Como puede observarse, la red de Poisson se ve más aleatoria y más homogénea que la red libre de escala[4]. En las redes de Poisson todos los nodos tienen más o menos el mismo número de conexiones. Algunos nodos estarán más conectados que otros, pero en promedio todos tienen la misma conectividad, es decir, las conexiones en una red de Poisson están distribuidas homogéneamente entre sus nodos. Por el contrario, la característica más importante de las redes libres de escala es su alta heterogeneidad, ya que existen nodos con muy pocas conexiones, nodos medianamente conectados y nodos extremadamente conectados. Los nodos altamente conectados se denominan los *núcleos* o *centros* de la red[5]. Con una red libre de escala uno no puede decir que todos los nodos tienen "más o menos" la misma conectividad. Por el contrario, hay nodos con una sóla conexión y también hay nodos con miles de conexiones.

Todavía no sabemos cuáles son los procesos que conducen a la formación de redes libres de escala. Sin embargo, en los últimos 8 a nos se han llevado a cabo avances sustanciales en la comprensión de estas redes. En las secciones que siguen veremos algunos de los formalismos matemáticos de crecimiento de redes que se han desarrollado para generar las tres topologías mencionadas

[4]Las redes con topología exponencial se parecen mucho a las de Poisson.

[5]En inglés se les llama *hubs*.

con anterioridad (Poisson, exponencial y libre de escala).

5. Redes de tipo Erdös-Rényi

Imaginemos un conjunto de N botones de pantalón distribuídos aleatoriamente sobre una mesa e inicialmente desconectados. Al tiempo $t = 0$ escogemos aleatoriamente una pareja de botones y los hilvanamos con un hilo. Después de haber enlazado a esta pareja, la dejamos sobre la mesa y escogemos aleatoriamente otra pareja para hilvanar. Podemos escoger botones que están conectados con otros botones, pero si la pareja que escogemos ya está conectada entre sí, la descartamos y escogemos otra pareja. Lo que no se vale es hilvanar más de una vez a la misma pareja de botones. Repetimos este proceso sucesivamente M veces, escogiendo aleatoriamente una pareja de botones cada vez. Al final del proceso habremos establecido M enlaces entre M parejas diferentes de botones, generando así una red de botones. Intuitivamente es claro que si M (el número total de enlaces) es peque no comparado con N (el número total de botones), entonces la red resultante estará desmembrada en varias islas peque nas. Dentro de cada isla los botones estarán hilvanados entre sí, pero estarán desconectados de las otras islas. Sin embargo, si M es muy grande comparado con N, terminaremos con casi todos los botones hilvanados unos con otros. Probablemente haya islas muy peque nas desconectadas de la red principal, pero seguramente la gran mayoría de botones formarán parte de una isla principal: la isla gigante.

Después de haber hilvanado M parejas en un conjunto total de N botones, ¿cuál es la distribución de conexiones $P(k)$ en la red resultante? Como veremos en un momento, la red que resulta de este proceso tiene una distribución $P(k)$ de Poisson. Pero antes de dar la prueba de este resultado, vale la pena mencionar que durante muchos a nos se pensó que este mecanismo de formación de redes en el cual parejas de nodos se enlazan aleatoriamente, era adecuado para describir el origen de ciertas redes sociales como las redes de amistades o las redes de contactos sexuales. Después de todo, las amistades o los contactos sexuales se dan por el encuentro casual y aleatorio de las personas que viven en una sociedad. Por lo tanto, era natural pensar que el mecanismo de "hilvanar parejas de botones escogidas al azar" reproducía lo que realmente ocurre en las redes sociales. Sí, era "natural"pensarlo, pero era incorrecto.

Calculemos ahora la probabilidad $P(k)$ para nuestra red de botones hil-

vanados. Para comenzar notemos que el número *total* N_p de parejas que se pueden formar en un conjunto de N botones es

$$N_p = \frac{1}{2}N(N-1)$$

Como enlazamos M parejas de botones, la probabilidad p_e de que una pareja arbitraria seleccionada al azar esté enlazada es

$$p_e = \frac{M}{N_p} = \frac{2M}{N(N-1)} \tag{4}$$

Ahora enfoquemos nuestra atención sobre un nodo particular v_j de la red, escogido al azar. El número total de parejas que podrían contener a v_j es $N-1$, ya que v_j se podría haber hilvanado con los $N-1$ nodos restantes de la red. Sin embargo, en los M enlaces que se llevaron a cabo, no necesariamente escogimos al nodo v_j todas las veces posibles que se podría haber escogido. Supongamos entonces que de las M parejas que se escogieron, el nodo v_j estaba solamente en k de ellas. La probabilidad de que v_j esté contenido en k parejas de las $N-1$ posibles es

$$P(k) = \binom{N-1}{k}(p_e)^k(1-p_e)^{N-1} \tag{5}$$

Esta es una distribución binomial para N y M finitas. Pero si consideramos ahora que la red es muy grande y tomamos el límite $N \to \infty$ y $M \to \infty$ de tal forma que la cantidad

$$z = \frac{2M}{N}$$

permanezca finita, entonces la distribución (5) se transforma en

$$P(k) = e^{-z}\frac{z^k}{k!} \tag{6}$$

lo cual es la distribución de Poisson con promedio z. En el apéndice A muestro los pasos algebráicos que conducen de la ecuación (5) a la ecuación (6).

6. Crecimiento de redes

En la sección anterior suponíamos que teníamos una población fija de N nodos y un número fijo M conexiones a nadidas aleatoriamente. Sin embargo,

en la realidad esto no ocurre, las redes no están fijas. Por el contrario, las redes complejas evolucionan y crecen en el tiempo a través de la adición simultánea tanto de conexiones como de nodos. Pensemos por ejemplo en la red internet. En octubre de 1969 la "redïnternet consistía de sólo dos computadoras, una en la Universidad de California en Los Angeles (UCLA) y la otra en el Instituto de Investigaciones de Stanford (SRI). El primer mensaje que se transmitieron estas computadoras fue "LOGWIN"[6]. A lo largo de los a nos más y más computadoras se sumaron a la red internet, y para el a no 2000, las primeras dos computadoras de la UCLA y el SRI que se dijeron "LOGWIN", se habían convertido ya en sistemas que conectaban a 170 países y a más de 300 millones de personas.

Así como la red internet nació siendo peque na y después creció con el paso del tiempo, también las redes metabólicas y las redes genéticas dentro de la célula, y muchas otras redes en la naturaleza, han crecido y evolucionado a lo largo del tiempo. Por lo tanto, es importante que nuestros modelos de formación de redes incorporen el hecho de que nuevos nodos y nuevas conexiones se pueden a nadir a la red. También debe tomarse en cuenta el que nodos y conexiones ya existentes pueden eliminarse.

En el modelo más simple de crecimiento de redes a nadimos un nuevo nodo en cada paso de tiempo. Este nuevo nodo puede conectarse con alguno de los nodos ya existentes. Cada uno de los nodos ya existentes pueden ser seleccionados para la conexión con una probabilidad $\Pi(k_i, t)$, siendo k_i la conectividad al tiempo t del i-ésimo nodo ya existente.

El proceso comienza con un único nodo inicial v_0 al tiempo $t = 0$ (ver la Fig. 5). Al tiempo $t = 1$ a nadimos un nuevo nodo v_1, que se conectará al único nodo ya existente v_0 con probabilidad 1. Después, al tiempo $t = 2$ a nadimos el nodo v_2 que se conectará con cualquiera de los nodos ya existentes v_1 y v_2 con la misma probabilidad 1/2. En este momento los nodos v_0, v_1 y v_2 ya no tienen todos la misma conectividad: alguno de ellos tendrá dos conexiones mientras que los otros dos nodos tendrán sólo una conexión. Al tiempo $t = 3$ a nadimos al nodo v_3, que se conectará con alguno de los nodos ya existentes v_0, v_1 o v_2 con una probabilidad que es función de sus conectividades k_0, k_1 y k_2. Continuando con este proceso, al tiempo $t + 1$ a nadimos al nodo v_{t+1} que se conectará a cualquiera de los nodos ya existentes v_0, v_1, ..., v_t, el cual será seleccionado con una probabilidad $\Pi(k_i, t)$, siendo k_i la conectividad al tiempo t del nodo seleccionado.

[6]LOGWIN es la concatenación de "log"(conectarse) y "win"(ganar).

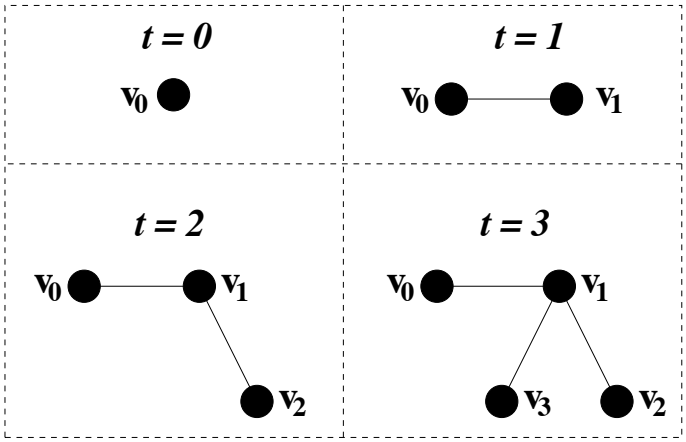


Figura 5: Crecimiento de redes. Al tiempo $t = 0$ hay un sólo nodo v_0. En cada paso de tiempo subsecuente a nadimos un nuevo nodo que se conectará a alguno de los nodos ya existentes con una probabilidad $\Pi(k, t)$ que depende de la conectividad k al tiempo t del nodo con el que se pretende establecer la conexión.

Denotemos por $P(k, t)$ a la probabilidad de que al tiempo t un nodo arbitrario de la red tenga k conexiones. Es claro que esta probabilidad depende del tiempo. Sin embargo, si continuamos a nadiendo nodos por un tiempo suficientemente largo esperamos que la función $P(k, t)$ alcance un estado estacionario independiente del tiempo. Esto no significa que la *red* alcance un estado estacionario. La red sigue creciendo mientras continuemos a nadiendo nodos. Es únicamente la distribución de conectividades $P(k, t)$ la que llega a un estado estacionario en el cual $P(k, t + 1) = P(k, t) = P(k)$[7].

Existen varios métodos para calcular la distribución de conectividades $P(k, t)$. En estas notas veremos el método de la *ecuación maestra*, pero existen por lo menos otros dos métodos distintos: el *método continuo*, inventado por Barabási, y el *método cinético*, introducido por Krapivsky, Redner, y Leyvraz [2]. Los tres métodos dan resultados equivalentes.

Para aquellos que no conocen el formalismo de la ecuación maestra, en el apéndice B doy una breve introducción al planteamiento de dicha ecuación, la cual es muy fácil de escribir pero, la mayoría de las veces, muy difícil de

[7] Análogamente, la distribución de estaturas en los habitantes del D.F. se ha mantenido estacionaria por muchos a nos aún cuando la población de chilangos no ha dejado de aumentar.

resolver.

Denotemos por $P(n,k,t)$ la probabilidad de que el nodo v_n tenga k conexiones al tiempo t. Notemos que esta probabilidad está asociada al nodo específico v_n. Sin embargo, podemos obtener la probabilidad $P(k,t)$ de que un nodo *arbitrario* tenga k conexiones al tiempo t a trevés de la siguiente ecuación:

$$P(k,t) = \frac{1}{N(t)} \sum_{n=0}^{t} P(n,k,t), \tag{7}$$

donde $N(t)$ es el número total de nodos de la red al tiempo t.

La ecuación que determina la evolución temporal de $P(n,k,t)$ se obtiene notando que en cada paso de tiempo hay dos contribuciones a dicha probabilidad:

1. Al tiempo t el nodo v_n tenía $k-1$ conexiones y fue seleccionado (con probabilidad $\Pi(k-1,t)$) para conectarse con el nuevo nodo a nadido a la red. Por lo tanto, al tiempo $t+1$ el nodo v_n tendrá k conexiones.

2. El nodo v_n ya tenía k conexiones al tiempo t y *no fue* seleccionado para conectarse con el nuevo nodo a nadido (lo cual ocurre con probabilidad $1-\Pi(k,t)$). Por lo tanto, al tiempo $t+1$ el nodo v_n seguirá teniendo k conexiones.

Tomando en cuenta estas dos contribuciones, la ecuación maestra que determina la evolución temporal de $P(n,k,t)$ es

$$P(n,k,t+1) = \underbrace{P(n,k-1,t)}_{\substack{k-1 \text{ conexiones} \\ \text{al tiempo } t}} \overbrace{\Pi(k-1,t)}^{\substack{v_n \text{ fue} \\ \text{seleccionado}}} + \underbrace{P(n,k,t)}_{\substack{k \text{ conexiones} \\ \text{al tiempo } t}} \overbrace{[1-\Pi(k,t)]}^{\substack{v_n \text{ no fue} \\ \text{seleccionado}}}. \tag{8}$$

Como el nodo v_n "nació" al tiempo $t=n$ con una sóla conexión, la condición inicial para resolver la ecuación anterior es

$$P(n,k,t)|_{t=n} = \delta_{k,1}. \tag{9}$$

Sumando sobre n en la Eq. (8) desde $n=0$ hasta $n=t$ y tomando en consideración la Eq. (7), obtenemos

$$\sum_{n=0}^{t} P(n,k,t+1) = N(t)\big\{P(k-1,t)\Pi(k-1,t) + P(k,t)\left[1-\Pi(k,t)\right]\big\} \tag{10}$$

Ahora bien, como en cada paso de tiempo a nadimos un nuevo nodo (comenzando con el nodo v_0 al tiempo $t = 0$), entonces $N(t) = t + 1$. Por lo tanto, la ecuación anterior puede escribirse como

$$\sum_{n=0}^{t} P(n,k,t+1) = (t+1)\big\{P(k-1,t)\Pi(k-1,t) + P(k,t)\left[1-\Pi(k,t)\right]\big\} \quad (11)$$

En el lado izquierdo de la ecuación anterior podemos completar la suma hasta $t + 1$ de la siguiente manera:

$$\begin{aligned}
\sum_{n=0}^{t} P(n,k,t+1) &= \sum_{n=0}^{\mathbf{t+1}} P(n,k,t+1) - P(t+1,k,t+1) \\
&= \sum_{n=0}^{\mathbf{t+1}} P(n,k,t+1) - \delta_{k,1} \\
&= N(t+1)P(k,t+1) - \delta_{k,1} \\
&= (t+2)P(k,t+1) - \delta_{k,1}
\end{aligned}$$

donde hemos utilizado el hecho de que $P(t+1,k,t+1) = \delta_{k,1}$, tal y como lo establece la Eq. (9). También utilizamos la Eq. (7) (evaluada en $t + 1$) y el hecho de que $N(t + 1) = t + 2$. Sustituyendo estos resultados en la Eq. (11) obtenemos

$$(t+2)P(k,t+1) - \delta_{k,1} = (t+1)\big\{P(k-1,t)\Pi(k-1,t) + P(k,t)\left[1-\Pi(k,t)\right]\big\} \quad (12)$$

Notemos que para poder resolver la Eq. (12) necesitamos conocer explícitamente la función $\Pi(k,t)$, que es la probabilidad de que un nodo ya existente con conectividad k sea seleccionado para conectarse con el nuevo nodo que se a nade al tiempo t. Como veremos en las siguientes secciones, diferentes formas de la probabilidad $\Pi(k,t)$ conducen a topologías diferentes.

7. Topología exponencial

En el caso en que cualquiera de los nodos ya existente pueda escogerse con la misma probabilidad para conectarse con el nuevo nodo a nadido, la probabilidad $\Pi(k,t)$ es independiente de k y queda dada por

$$\Pi(k,t) = \frac{1}{N(t)} = \frac{1}{t+1}$$

donde $N(t) = t + 1$ es el número total de nodos al tiempo t. Sustituyendo esta forma de $\Pi(k)$ en la Eq. (12) obtenemos

$$(t+2)P(k,t+1) - \delta_{k,q} = P(k-1,t) + tP(k,t)$$

Después de un tiempo muy largo, la función $P(k,t)$ alcanza un estado estacionario en el que $P(k,t+1) = P(k,t) = P(k)$. En dicho estado estacionario la última ecuación se transforma en

$$P(k) = \frac{1}{2}\big(P(k-1) + \delta_{k,1}\big). \tag{13}$$

Como todos los nodos de la red tienen por lo menos una conexión, es claro que $P(0) = 0$. Con esta condición inicial, la ecuación de recurrencia anterior tiene la solución

$$P(k) = 2^{-k} \tag{14}$$

que no es más que la distribución exponencial. Es importante se nalar que esta distribución aparece en el contexto del crecimiento de redes como el resultado de lo que se llama *enlace igualitario*, es decir, en una situación en la cual cada nodo nuevo que se a nade a la red se puede enlazar a cualquiera de los nodos ya existentes con la misma probabilidad. En este sentido el nuevo enlace que se forma es igualitario, porque no discrimina entre los nodos ya existentes.

8. Topología libre de escala

En la vida real las conexiones entre diferentes nodos no se dan de manera igualitaria. Por ejemplo, si tenemos una computadora nueva y queremos conectarla a internet, no vamos a contratar el servicio de internet de alguna compa nia elegida al azar, sino que buscaremos la compa nia que ofrezca el mejor servicio y al mejor precio, y probablemente será esta compa nia la que tenga más clientes. En una escuela los varones no buscan a su pareja al azar, sino que buscarán salir con la chica más bonita, o tal vez con la más inteligente, y será esta muchacha la que tenga más pretendientes. Por esta razón, Barabási inventó el concepto de *enlace preferencial* en el cual los nuevos nodos que se a naden a la red se conectarán preferentemente con los nodos ya existentes que tengan el mayor número de conexiones. Intuitivamente podemos pensar que el enlace preferencial consiste en que uno siempre trata de

estar conectado con los nodos más "populares", es decir, con los nodos de mayor conectividad.

Para incorporar este comportamiento, Barabási sugirió que la probabilidad de enlace $\Pi(k,t)$ debe tomar la forma

$$\Pi(k,t) = \left(\sum_{n=0}^{N(t)} k_n\right)^{-1} k \tag{15}$$

donde k_n es la conectividad del n-ésimo nodo ya existente al tiempo t. El factor $\left(\sum_{n=0}^{N(t)} k_n\right)^{-1}$ es simplemente para garantizar que la probabilidad $\Pi(k,t)$ esté normalizada. Al hacer que $\Pi(k,t)$ sea proporcional a k, como lo propuso Barabási, tenemos enlace preferencial, ya que de esta forma, entre más grande sea la conectividad k de un nodo, mayor será la probabilidad de conectarse con él. Como en cada paso de tiempo a nadimos un nuevo nodo con una conexión, comenzando con cero conexiones al tiempo $t = 0$, entonces para cualquier tiempo $t > 0$ se tiene

$$\sum_{n=0}^{N(t)} k_n = 2t,$$

y por lo tanto

$$\Pi(k,t) = \frac{k}{2t} \tag{16}$$

Substituyendo este resultado en la Eq. (12) obtenemos

$$(t+2)P(k,t+1) - \delta_{k,q} = \frac{t+1}{2t}\left\{(k-1)P(k-1,t) + [2t-k]\,P(k,t)\right\}$$

En el límite $t \to \infty$ el sistema alcanza el estado estacionario. Por lo tanto, la distribución de conexiones estacionaria $P(k)$ se obtiene de la ecuación anterior tomando el límite $t \to \infty$, lo que conduce a

$$P(k) + \frac{1}{2}\big[kP(k) - (k-1)P(k-1)\big] = \delta_{k,1} \tag{17}$$

Como todos los nodos tienen por lo menos una conexión, para resolver la ecuación (17) utilizamos la condición inicial $P(0) = 0$, lo cual nos da la solución

$$P(k) = \frac{4}{k(k+1)(k+2)} \tag{18}$$

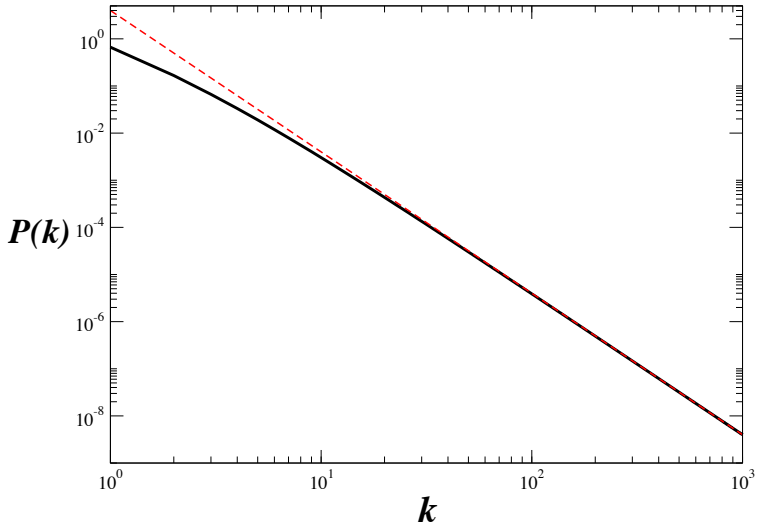


Figura 6: Gráfica log-log de la distribución de conectividades $P(k)$ dada en la ecuación (18). La línea punteada roja es la gráfica de la ley de potencias "perfecta" $f(k) = 4k^{-3}$. Nota que $P(k) \approx 4k^{-3}$ para valores grandes de k.

Aun cuando esta no es una ley de potencias "perfecta" (ver la Fig. 6), para valores grandes de k la distribución $P(k)$ se comporta como $P(k) \sim k^{-3}$. Por un lado, este es un resultado muy interesante: el enlace preferencial genera distribuciones de conectividades con colas libres de escala. Son estas colas largas las responsables de que existan elementos altamente conectados. Por otro lado, este resultado también es desalentador ya que el proceso de enlace preferencial *siempre* da el exponente $\gamma = 3$, el cual no se encuentra frecuentemente en la naturaleza (ver los exponentes en los cuadros 3 y 4). En otras palabras, aun cuando el método de enlace preferencial nos da *una* ley de potencias, es incapaz de reproducir la gran variedad de exponentes encontrados en la naturaleza. Por esta razón, se han propuesto diferentes formas de la función $\Pi(k,t)$ que corresponden a diferentes tipos de enlace preferencial. Por ejemplo, Krapivsky, Redner, y Leyvraz propusieron una función de enlace $\Pi(k,t)$ *no lineal* (ver la referencia [6]), de la siguiente forma

$$\Pi(k,t) = \left(\sum_{n=0}^{N(t)} k_n^{\alpha} \right)^{-1} k^{\alpha} \tag{19}$$

donde α es un exponente arbitrario. Desafortunadamente, sólo para $\alpha = 1$ se tiene que la distribución $P(k)$ tiene una cola libre de escala. El trabajo de Krapivsky, Redner, y Leyvraz mostró que la naturaleza libre de escala de la red queda destruida cuando el enlace preferencial obedece una regla no lineal como la dada en la Eq. (19) cuando $\alpha \neq 1$. Es importante mencionar que todavía no tenemos modelos de crecimiento de redes que generen todos los exponentes listados en los cuadros 3 y 4.

9. ¿Nodos que se hacen viejos?

Otro problema con el modelo de enlace preferencial propuesto por Barabási es que predice que los nodos más viejos, es decir lo que se a nadieron primero a la red, son los que eventualmente adquirirán el mayor número de conexiones. En otras palabras, en una red libre de escala podríamos identificar facilmente a los nodos más viejos: son aquellos altamente conectados. Y entre más viejo sea un nodo, mayor será su conectividad.

Para ver que esto efectivamente es un resultado del modelo de enlace preferencial, consideremos un nodo particular v_n de la red. Su conectividad k_n va a cambiar a una tasa que es proporcional a la probabilidad $\Pi(k_n, t)$ de que este nodo adquiera más conexiones. Es decir,

$$\frac{dk_n}{dt} = \Pi(k_n, t)$$

Para el caso particular en el que $\Pi(k, t)$ está dada como en la Eq. (16), tenemos

$$\frac{dk_n}{dt} = \frac{k_n}{2t}$$

Como el nodo v_n nació al tiempo $t = n$ con una conexión, la condición inicial para la ecuación anterior es $k_n(n) = 1$, lo cual conduce a la solución

$$k_n(t) = \left(\frac{t}{n}\right)^{1/2} \tag{20}$$

De la ecuación anterior, es claro que a cualquier tiempo t se cumple la siguiente desigualdad

$$k_1(t) > k_2(t) > k_3(t) > \cdots > k_n(t) \tag{21}$$

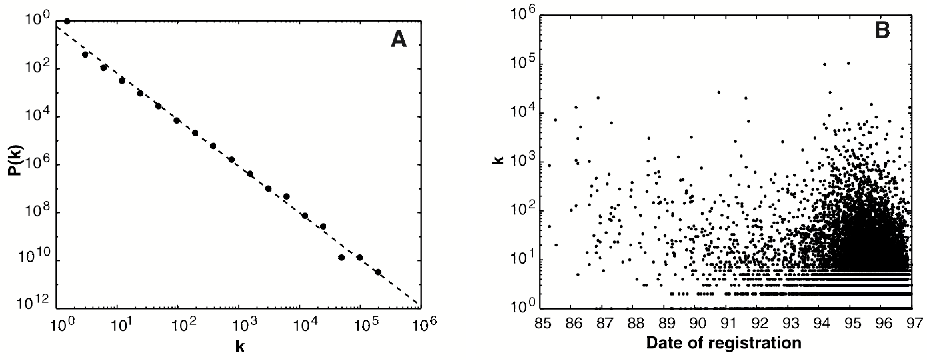


Figura 7: (A) Gráfica log-log de la distribución de conectividades $P(k)$ para una red WWW consistente en 260,000 páginas. (B) Conectividad k de los nodos de la red como función de su edad. Las figuras fueron tomadas (sin permiso) de la referencia [7].

lo cual significa que los nodos que nacieron primero tendrán, en promedio, conectividades mayores que los nodos que llegaron después.

Sin embargo, este comportamiento predicho por el modelo del enlace preferencial, en el que los nodos más viejos son los más conectados, no siempre se observa en la naturaleza. Por ejemplo, Lada A. Adamic y Bernardo Huberman estudiaron una red que consistía en 260,000 páginas www [7], considerando que dos páginas estaban conectadas si una contenía un hipervínculo a la otra. Adamic y Huberman encontraron que, aún cuando esta red tiene topología libre de escala con un exponente $\gamma \sim 2$ (ver la Fig. 7(a)), no existe correlación entre la conectividad de los nodos y su edad. La Fig. 7(b) muestra la gráfica de la conectividad de los nodos como función de su edad. Como puede observarse, estas dos cantidades no parecen estar relacionadas.

Probablemente el contra ejemplo más contundente que contradice los resultados de las Eqs. (20) y (21) es Google, el robot buscador de la red WWW. Recordemos que Altavista y Yahoo ya existían y estaban bien establecidos en 1998, a no en que apareció Google. Rápidamente Google tomó la delantera, convirtiéndose en el robot buscador más popular en el mundo informático. ¿A qué se debió este éxito repentino de Google? Probablemente a que estaba mejor dise nado, era más rápido y más eficiente que sus competidores Altavista y Yahoo. En otras palabras, Google nació con características que lo hacían un robot "mejor adaptado" que sus competidores.

Esta observación hizo que Barabási propusiera el concepto de *adaptabi-*

lidad en el contexto de las redes complejas. Así, cada uno de los nodos v_n, además de tener una conectividad k_n, también tenia un parámetro de adaptabilidad asociado w_n. Este parámetro es una medida de qué tan bien estaba adaptado el nodo v_n a su entorno: entre más grande es el valor de w_n, mayor es el grado de adaptabilidad de v_n. Específicamente, lo que Barabási hizo fue proponer una función de enlace preferencial $\Pi(k, w, t)$ que, además de ser proporcional a la conectividad k de los nodos ya existentes, también es proporcional a su adaptabilidad w (ver la referencia [8]). De esta forma, la probabilidad de que un nuevo nodo a nadido a la red se conecte con el nodo v_n ya existente, cuya conectividad es k_n y cuya adaptabilidad es w_n, queda dada por

$$\Pi(k_n, w_n, t) = Cw_nk_n$$

donde C es una constante de normalización. Las adaptabilidades w_n asociadas a los nodos de la red son variables aleatorias que se escogen de una distribución $A(w)$. Cada nodo nace con su propia adaptabilidad, la cual no cambia en el tiempo.

Un resultado sorprendente del modelo de enlace preferencial con adaptabilidad es que puede ser mapeado exactamente a un gas de Bose-Einstein. En este mapeo cada nodo de la red representa un nivel energético del gas, mientras que las conectividades de los nodos representan los números de ocupación de los respectivos niveles de energía. El mapeo entre el modelo de red y el gas de Bose-Einstein es más que una simple matáfora. Es un mapeo matemático preciso y, bajo ciertas condiciones, la red en crecimiento puede presentar "condensación de Bose-Einstein", lo cual significa que repentinamente un nodo puede adquirir la mayoría de las conexiones de la red, independientemente de su edad (como sucedió con Google). Los que estén interesados podrán encontrar los detalles de este trabajo en la referencia [8]. Hasta donde yo sé, no existe acuerdo cuantitativo entre los resultados del modelo de enlace preferencial con adaptabilidad, y los datos experimentales que se tienen de redes reales. Sin embargo, la condensación de Bose-Einstein predicha en este modelo podría ser el primer paso para entender mejor el "fenómeno Google".

10. La isla gigante

Regresemos al modelo de los botones descrito en la sección 5. Inicialmente todos los botones están sobre la mesa desconectados, pero conforme

vamos hilvanando parejas de botones, se comienzan a formar islas de botones conectados. Al princípio las islas son peque nas, pero al ir a nadiendo más y más conexiones entre los botones, las islas crecen y se conectan entre sí. Eventualmente se formará una isla gigante, es decir, una isla que es mucho más grande que todas las demás.

El tama no de la isla gigante es importante en el estudio de la propagación de epidemias en una sociedad, en donde en lugar de tener botones hilvanados, tenemos personas que se contagian unas a otras. Las islas consisten en los grupos de personas que están infectadas, y la enfermedad se convierte en una epidemia cuando se forma una isla gigante que abarca a la mayoría de la sociedad.

La pregunta importante es: dado un conjunto de elementos, ¿cuántos enlaces (contagios) tienen que establecerse para que se forme la isla gigante? Esta pregunta fue contestada por Erdös y Rényi, quienes fueron los primeros en mostrar la existencia de una transición de fase en la teoría de redes. Dicha transición de fase consiste precisamente en la formación de la isla gigante. Aunque su trabajo original lo llevaron a cabo para redes con topología de Poisson, es fácil generalizar el análisis que hicieron para extenderlo a redes con topologías arbitrarias.

Para calcular el tama no de la isla gigante utilizaremos un razonamiento de consistencia. Sea q la probabilidad de que un nodo v_n escogido aleatoriamente *no pertenezca* a la isla gigante. Supongamos que v_n tiene k vecinos. Claramente, v_n no pertenece a la isla gigante si y sólo si ninguno de sus k vecinos tampoco pertenece a la isla gigante. Por lo tanto, la probabilidad q debe satisfacer la ecuación de consistencia

$$q = \sum_k P(k) q^k \tag{22}$$

El lado izquierdo de esta ecuación es simplemente la probabilidad q de que v_n no pertenezca a la isla gigante. El lado derecho es la probabilidad $P(k)$ de que v_n tenga k vecinos, multiplicada por la probabilidad q^k de que ninguno de estos k vecinos pertenezca a la isla gigante. La suma sobre k toma en cuenta todos los posibles vecinos que v_n pudiera llegar a tener.

En el caso en que $P(k)$ es una distribución de Poisson, la Eq. (22) se transforma en

$$q = \sum_{k=0}^{\infty} e^{-z} \frac{(zq)^k}{k!} = e^{z(q-1)}$$

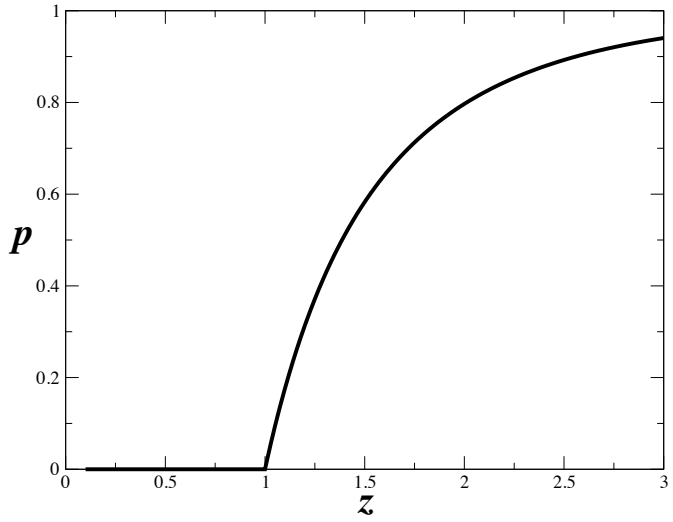


Figura 8: Probabilidad p de pertenecer a la isla gigante como función de la conectividad promedio z de la red, pare el caso de la topología de Poisson.

Denotando como $p = 1 - q$ la probabilidad de que un nodo arbitrario *si* pertenezca a la isla gigante, obtenemos

$$p = 1 - e^{-zp} \tag{23}$$

La ecuación anterior es trascendental y no puede resolverse de forma analítica. Sin embargo, es fácil resolverla de forma numérica y encontrar p para cada valor de z. La Fig. 8 muestra p como función de z. Notemos que para $z < 1$ la probabilidad de pertenecer a la isla gigante es cero. En otras palabras, no hay isla gigante para $z < 1$. Conforme aumentamos el valor de z (lo cual es equivalente a a nadir más conexiones a la red), la isla gigante comienza a formarse y justo en $z = 1$ aparece. Este es un resultado interesante: se necesita, en promedio, sólo una conexión por nodo para que aparezca la isla gigante. Si seguimos aumentando el valor de z, el tama no de la isla gigante aumenta más y más, como se puede observar en la Fig. 8, la cual muestra el comportamiento característico de una transición de fase de segundo orden que ocurre en $z = 1$.

Consideremos ahora el caso de las redes con topología libre de escala, para las que la distribución $P(k)$ toma la forma

$$P(k) = \frac{1}{\zeta(\gamma)} k^{-\gamma}$$

donde $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$ es la función zeta de Riemman la cual, en este caso simplemente juega el papel de la constante de normalización. Para esta distribución la ecuación de consistencia Eq. (22) se convierte en

$$\begin{aligned} q &= \frac{1}{\zeta(\gamma)} \sum_{k=1}^{\infty} \frac{q^k}{k^\gamma} \\ &= \frac{1}{\zeta(\gamma)} \mathrm{Li}_\gamma(q) \end{aligned}$$

donde $\mathrm{Li}_\gamma(q) = \sum_{k=1}^{\infty} k^{-\gamma} q^k$ es la función poli-logarítmica de orden γ. En términos de la probabilidad $p = 1 - q$, la ecuación anterior queda como

$$p = 1 - \frac{1}{\zeta(\gamma)} \mathrm{Li}_\gamma(1 - p)$$

Como $\mathrm{Li}_\gamma(0) = 0$, es claro $p = 1$ siempre es olución de la ecuación anterior. Por lo tanto, en una red libre de escala *la componente gigante siempre existe y la probabilidad de pertenecer a ella es* $p = 1$. Este interesante resultado es una consecuencia de la existencia de los elementos altamente conectados en redes libres de escala. Dichos elementos mantienen a toda la red conectada, impidiendo que se fracture en peque nas islas.

El resultado anterior puede tener implicaciones importantes en la propagación de epidemias en una sociedad, en la que podemos considerar que la isla gigante es el conjunto de personas que se han contagiado de alguna enfermedad. Los resultados presentados en esta sección muestran que en redes con topología de Poisson siempre podemos vacunar a un número suficiente de personas para garantizar que la conectividad promedio z de la red de individuos infectados se mantenga inferior a 1. En tal caso podemos detener la epidemia, ya que no habrá isla gigante, sólo peque nas islitas de personas infectadas.

Sin embargo, las redes sociales no tienen topología de Poisson, sino que exhiben topologías libres de escala. Para tales topologías siempre existe una isla gigante, y la probabilidad de pertenecer a ella es $p = 1$. Por lo tanto, en redes libres de escala las enfermedades tarde o temprano se propagan a todas las personas que no hayan sido vacunadas. Es decir, la única forma de detener la propagación de una epidemia en redes libres de escala es vacunando a *todas* las personas de la sociedad. Los efectos aterradores del resultado

anterior pueden observarse claramente en áfrica, donde comunidades enteras han desaparecido a causa del SIDA.

Más información sobre la propagación de epidemias en redes sociales puede encontrarse en las referencias [9, 10].

11. Procesos dinámicos en redes

Hasta ahora hemos visto las propiedades estructurales o topológicas de las redes complejas. Estas propiedades nos dicen cómo están conectados los elementos de la red y su estudio nos muestra de qué forma procesos aleatorios y aparentemente independientes pueden dar lugar a estructuras complejas e inesperadas (como la topología libre de escala). No obstante, sabemos que una vez que los elementos de una red están conectados, éstos pueden comenzar a interactuar generando procesos dinámicos que se propagan a través de toda la red. Consideremos, por ejemplo, las neuronas en el cerebro, las cuales están conectadas unas con otras por medio de axones y dendritas. A través de estas conexiones las neuronas interaccionan enviándose impulsos eléctricos que se propagan por todo el cerebro y que generan patrones de actividad eléctrica en las diferentes capas cerebrales responsables de la visión, el olfato, el tacto, etc. Son precisamente estos patrones de actividad eléctrica neuronal, asociados con nuestros sentidos y con nuestra consciencia, lo que llamamos "procesos dinámicos" en el cerebro.

Otro ejemplo interesante de un proceso dinámico en redes es la formación de opiniones en una sociedad. En este caso, la red social la componen los individuos de la sociedad ligados entre sí por medio de interacciones de amistad, camaradería, parentesco, etc. Los individuos intercambian opiniones con sus amigos, colegas, compañeros o familiares respecto a un determinado tema, y es gracias a este intercambio de opiniones que los individuos se van formando su propia opinión. Las opiniones se propagan a través de toda la sociedad, y eventualmente en la sociedad existirá una opinión mayoritaria respecto a dicho tema. El libre intercambio de opiniones es la base de la democracia en una sociedad moderna, en la que se supone que los individuos eligen, por mayoría, a sus gobernantes. Este es precisamente el tema que abordaremos en este capítulo: la formación de una opinión mayoritaria en una sociedad.

11.1. El modelo de votantes

Supongamos que en una sociedad existe un tema respecto al cual se pueden tener sólo dos opiniones. Por ejemplo, el tema del aborto, respecto al cual se puede estar a favor o en contra. O la privatización de la industria eléctrica, respecto de la cual cada uno de nosotros podemos estar a favor o en contra. (No valen opiniones indeterminadas como el típico "no sé"). Cada individuo tiene su propia opinión respecto a dicho tema. Representemos la opinión de un individuo con la variable σ, la cual puede tomar sólo dos valores: $\sigma = +1$ si el individuo está a favor, y $\sigma = -1$ si el individuo está en contra. Como en una sociedad existen N individuos, entonces tenemos un conjunto de N variables $\sigma_1, \sigma_2, \ldots, \sigma_N$, cada una representando la opinión de un individuo en la sociedad. De esta forma, $\sigma_n = +1$ si el n-ésimo individuo está a favor, mientras que $\sigma_n = -1$ si el n-ésimo individuo está en contra.

Lo interesante es que cada individuo intercambia opiniones con sus amigos y, en base a dicho intercambio, se forma su propia opinión. Supongamos, por tanto, que la opinión de cada individuo está influenciada por sus amigos (o conocidos, o familiares, etc.). Sea k_n el número de personas que tienen influencia sobre la opinión del n-ésimo individuo, y sean $\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}$ las k_n personas que tienen influencia sobre el individuo σ_n. Estríctamente hablando, las variables σ_n representan las *opiniones* de las personas, no a las personas mismas. Sin embargo, utilizaremos un lenguaje simple y diremos indistintamente que σ_n es la opinión del n-ésimo individuo, o bien σ_n es el n-ésimo individuo.

Decíamos entonces que cada individuo σ_n recibe opiniones de otros k_n individuos en la sociedad, a los cuales denotamos como $\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}$. Llamaremos a este conjunto *los inputs* de σ_n. En otras palabras, los inputs de σ_n son todas las personas que tienen influencia en su opinión, ya sean amigos, conocidos, parientes, maestros, etc. Cada individuo tiene su conjunto particular de inputs, lo cual significa que dos individuos diferentes σ_n y σ_m tendrán, en general, inputs distintos (aunque algunos de sus inputs pueden ser comunes). La red social se forma estableciendo quién recibe opiniones de quién, es decir, estableciendo para cada individuo el conjunto de personas que influencian su opinión. Notemos que esta red es dirigida, ya que si σ_i es un input de σ_j, entonces no necesariamente σ_j será un input de σ_i.

Una vez que sabemos cuáles son los inputs de cada individuo en la sociedad, tenemos que establecer la regla dinámica a través de la cual los individuos influyen en las opiniones de los demás. Dicha regla está basada en

la siguiente observación sencilla: *en general, un individuo tiende a ser de la misma opinión que la mayoría de sus amigos (inputs)*. Si lo pensamos un poco nos daremos cuenta de que esta regla es bastante cierta. En general, tendemos a juntarnos con personas que tienen (más o menos) la misma forma de pensar que nosotros. El objetivo de este capítulo es ver como esta sencilla regla de interacción, en la que un individuo "tiende" a ser de la misma opinión que la mayoría de sus inputs, puede generar estados de orden colectivo en toda la sociedad.

Establezcamos matemáticamente la regla dinámica. Sean σ_{n_1}, $\sigma_{n_2}, \ldots,$ $\sigma_{n_{k_n}}$ las opiniones de los k_n inputs de σ_n. Estas opiniones cambian a lo largo del tiempo, y supondremos que el tiempo lo medimos en unidades discretas que pueden ser días o semanas. Por ejemplo, cada semana hacemos una encuesta para monitorear las opiniones de las personas en la sociedad. Nuestra primer regla dinámica consiste en que el valor de la variable σ_n al tiempo $t+1$ está determinado por el valor de sus inputs al tiempo t de acuerdo con la siguiente ecuación:

$$\sigma_n(t+1) = \text{Signo}\left[\frac{1}{k_n}\sum_{j=0}^{k_n} w_{n_j}\sigma_{n_j}(t)\right], \tag{24}$$

donde la función Signo$[x]$ se define como

$$\text{Signo}[x] = \begin{cases} +1 & \text{si} \quad x \geq 0 \\ -1 & \text{si} \quad x < 0 \end{cases}$$

Por otro lado, la cantidad $\dfrac{1}{k_n}\displaystyle\sum_{j=0}^{k_n} w_{n_j}\sigma_{n_j}(t)$ que aparece dentro de la función Signo es el promedio ponderado de las opiniones de los inputs de σ_n al tiempo t. Los factores de w_{n_j} son números reales que representan el peso de las opiniones de los inputs. Estos factores toman en cuenta el hecho de que no todos los amigos de un individuo tienen la misma influencia sobre su opinión. Habrá algunas personas a las que dicho individuo respetará más que a otras. Por lo tanto, en esta suma ponderada, entre más grande sea el valor de algún w_{n_j}, más importancia tendrá el correspondiente input σ_{n_j} para determinar el valor de σ_n.

La Eq. (24) nos dice entonces que $\sigma_n = +1$ si la mayoría (ponderada) de sus inputs tienen opinión $+1$, mientras que $\sigma_n = -1$ si la mayoría (ponderada) de sus inputs tienen opinión -1. A cada instante de tiempo, todos los

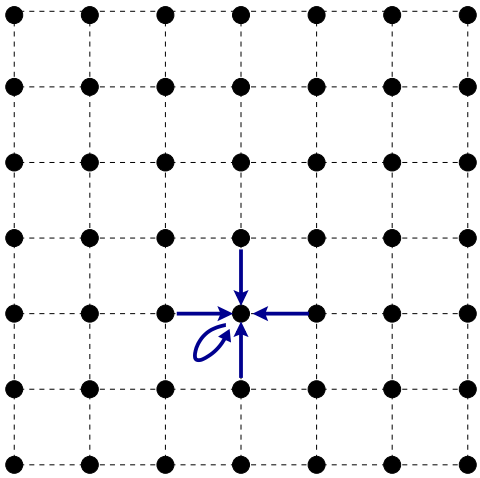


Figura 9: Los individuos de la sociedad están representados por los círculos negros. Los colocamos en una red cuadrada de tal manera que cada individuo tiene 5 inputs, sus cuatro primeros vecinos y él mismo. Las flechas indican los inputs del individuo particular seleccionado. La misma configuración se tiene para todos los individuos.

individuos de la red actualizan su valor de acuerdo con la Eq. (24). Ilustremos la dinámica generada por esta ecuación en un caso sencillo.

Supongamos que colocamos a los individuos en una red cuadrada de tal modo que cada individuo recibe inputs de sus cuatro primeros vecinos y de sí mismo (porque su opinión también cuenta), como se indica en la Fig. 9. Como todos los individuos tienen 5 inputs, entonces $k_n = 5$ para todos. Además, vamos a suponer que todos los pesos w_{n_j} son iguales a 1, es decir, los inputs de un individuo tienen todos la misma influencia sobre su opinión. En esta caso, la dinámica generada por la Eq. (24) es muy aburrida. La Fig. 10 muestra un sistema con 10000 individuos acomodados en una red cuadrada de 100×100. Cada cuadrito es un individuo. Aquellos con opinión -1 se pintan con color negro, mientras que los individuos con opinión +1 se pintan en un color más claro. Al tiempo $t = 0$ las opiniones +1 y -1 están distribuidas al azar en la sociedad de tal forma que aproximadamente la mitad de los individuos tienen opinión +1 y la otra mitad tienen opinión -1. Después de iterar la Eq. (24) cien veces ($t = 100$), el sistema alcanza un estado estacionario en el que se forman *grupos de opinión*, es decir, individuos con la misma opinión tienden a agregarse en cúmulos. Aunque dejemos correr el tiempo más y más, la configuración mostrada en el panel derecho de la

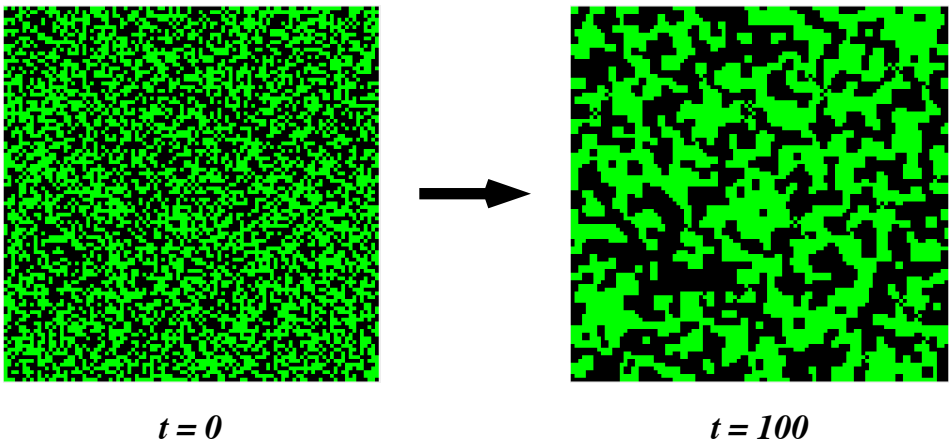


Figura 10: Cada cuadrito representa un individuo, negro si el individuo tiene opinión -1 y claro si el individuo tiene opinión +1. El cuadro grande de la izquierda es la condición inicial con 50 % de opiniones positivas y 50 % de opiniones negativas repartidas aleatoriamente en la sociedad. El cuadro de la derecha es el estado al que llega el sistema después de 100 pasos de tiempo iterando la Eq. (24). Aun cuando se forman grupos de opinión, los porcentajes de opiniones positivas y negativas no cambian con el tiempo y permanecen iguales a lo que se tenía en la condición inicial.

Fig. 10 ya no cambia. Es interesante notar que, aunque se formaron grupos de opinión, los porcentajes de opiniones positivas y negativas siguen siendo los mismos que al principio, 50 % y 50 % respectivamente. Es decir, la Eq. (24) *no genera una opinión mayoritaria en la sociedad*, simplemente preserva los porcentajes iniciales de opiniones.

Este ejemplo muestra que la vida sería muy sencilla si nuestras opiniones estuvieran regidas por la Eq. (24), es decir, si todo mundo hiciéramos lo que la mayoría de las personas nos dicen que hagamos. Sin embargo, somos necios en el sentido de que, incluso cuando la mayoría de nuestros amigos tengan una opinión respecto a un tema, con una determinada probabilidad nosotros podemos tener la opinión contraria. Por lo tanto, modificaremos la Eq. (24) para introducir el *libre albedrío* de las personas. Esto lo haremos suponiendo que cada individuo en la sociedad adquiere la opinión contraria a la de la mayoría de sus inputs con probabilidad η, o bien, adquiere la opinión de la mayoría de sus inputs con probabilidad $1-\eta$. Es decir, la regla dinámica es

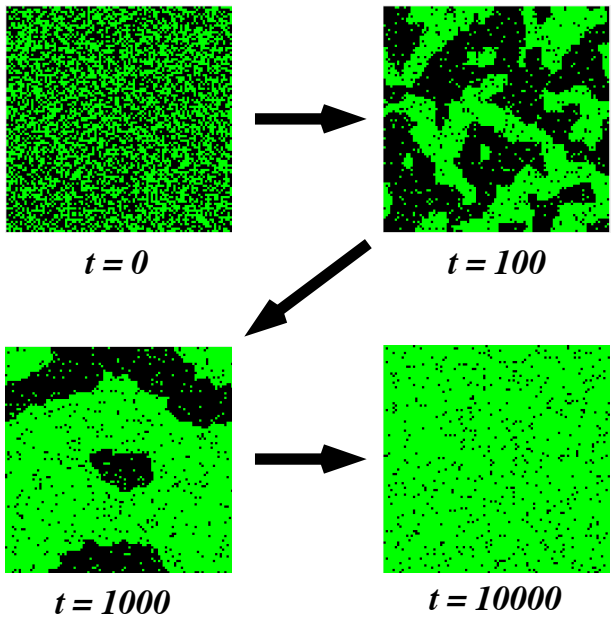


Figura 11: En este caso la dinámica del sistema está regida por la Eq. (25) con una probabilidad $\eta = 0.05$ de violar la regla de la mayoría. El primer cuadro al tiempo $t = 0$ es la condición inicial, con 50 % de opiniones positivas y 50 % de opiniones negativas repartidas aleatoriamente en la sociedad. Vemos que conforme pasa el tiempo se va formando una opinión mayoritaria en la sociedad, hasta que claramente al tiempo $t = 10000$ la gran mayoría de los individuos en la sociedad tienen opinión $+1$.

ahora

$$\sigma_n(t+1) = \begin{cases} \text{Signo}\left[\sum_{j=0}^{k_n} w_{n_j}\sigma_{n_j}(t)\right] & \text{con probabilidad} \quad 1-\eta \\ -\text{Signo}\left[\sum_{j=0}^{k_n} w_{n_j}\sigma_{n_j}(t)\right] & \text{con probabilidad} \quad \eta \end{cases} \tag{25}$$

El parámetro η es entonces la probabilidad de que cada individuo vaya en contra de la opinión mayoritaria de sus inputs.

La Fig. 11 muestra la dinámica generada por la Eq. (25) para el caso de la red cuadrada mostrada en la Fig. 9 (otravez, con todos los pesos $w_{n_j} = 1$). Escogimos para este ejemplo una probabilidad $\eta = 0.05$ muy pequeña de ir en contra de la opinión mayoritaria de los inputs. Como muestra la Fig. 11, comenzamos al tiempo $t = 0$ con una condición inicial con la mitad de opiniones positivas y la otra mitad de las opiniones negativas, distribuidas alea-

toriamente en la sociedad. Conforme pasa el tiempo se forman los grupos de opinión, es decir, individuos con la misma opinión tienden a agregarse (como en el panel correspondiente al tiempo $t = 100$). Sin embargo, al transcurrir el tiempo notamos que ocurre algo verdaderamente importante: *emerge una opinión mayoritaria en la sociedad* de tal forma que al tiempo $t = 10000$ la gran mayoría de los individuos tienen la misma opinión (que en este ejemplo es $+1$). Vemos entonces que el *ruido* que introducen las personas que hacen lo contrario de lo que dicen la mayoría de sus inputs (los famosísimos "contreras", o "anarquistas", o "revoltosos", etc.), representados por el parámetro η, son fundamentales para la emergencia de orden en la sociedad. Notemos que la Eq. (24), la cual no genera una opinión mayoritaria, es un caso particular de la Eq. (25) con $\eta = 0$. Por lo tanto, si en nuestra sociedad no hay "revoltosos", si todo el mundo hace lo que dicen los demás, no se genera orden. Es fundamental la presencia de unos cuantos "revoltosos" para la emergencia de orden, es decir, para la formación de una *opinión común* en toda la sociedad. Esto es sorprendente porque generalmente tendemos a pensar que el ruido (los "revoltosos") destruye el orden, y lo consideramos como algo no deseado. Sin embargo, este ejemplo muestra que el ruido muchas veces es necesario para tener orden en un sistema. Sorprendente, ¿no?

11.2. Transición de fase

En el ejemplo mostrado en la Fig. 11 el valor del parámetro η que determina la cantidad de ruido presente en el sistema es muy pequeño: $\eta = 0.05$ (esto quiere decir que hay muy pocos revoltosos). En tal caso la presencia del ruido es fundamental para generar un orden global, es decir, una opinión mayoritaria en toda la sociedad. Sin embargo, mucho ruido tampoco es bueno porque entonces el orden se destruye. Esto se ilustra en la Fig. 12, la cual se generó utilizando el valor $\eta = 0.15$. Como podemos ver, el sistema nunca se ordena. Nunca emerge una opinión global común en la sociedad. (Incluso si comenzaramos con una condición inicial totalmente ordenada en la que todos los individuos tienen la misma opinión, la presencia de muchos revoltosos destruiría este orden.) Unos cuantos revoltosos ayudan a generar orden, pero muchos lo destruyen.

Para ver qué tanto ruido se necesita para destruir el orden en la sociedad vamos a definir un parámetro que mida la cantidad de orden. A este parámetro se le llama (no muy imaginativamente) *el parámetro de orden*, y se define

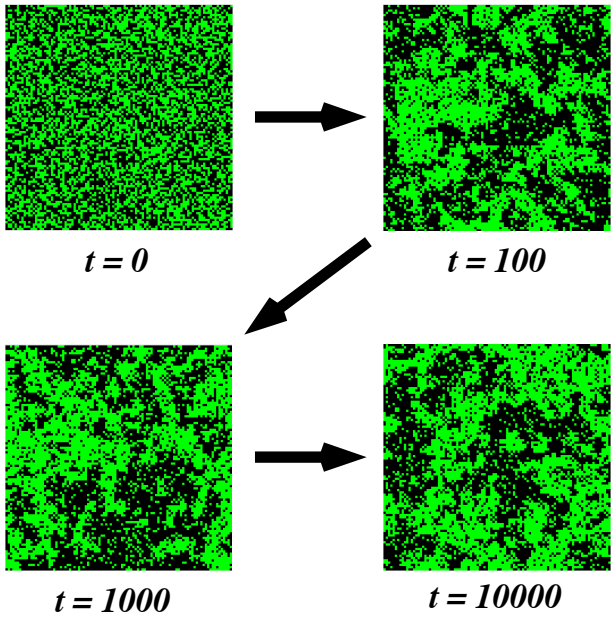


Figura 12: Ejemplo de la dinámica generada por la Eq. (25) con un valor relativamente alto de ruido: $\eta = 0.15$. En este caso nunca se genera orden en el sistema ya que el ruido es tan alto que lo destruye.

como

$$\psi(t) = \frac{1}{N}\left|\sum_{n=0}^{N}\sigma_n(t)\right|. \tag{26}$$

De acuerdo con esta definición, $\psi(t)$ es el valor absoluto de la opinión promedio de la sociedad al tiempo t (la suma se hace sobre todos los individuos en la sociedad). Si aproximadamente la mitad de las opiniones son positivas y la otra mitad negativas, entonces $\psi(t) \approx 0$ y no hay orden en el sistema. Por el contrario, si casi todos los individuos tienen la misma opinión (ya sea positiva o negativa), entonces $\psi(t) \approx 1$. En tal caso decimos que el sistema está muy ordenado.

Cuando $t \to \infty$, el sistema alcanza un estado estacionario en el cual el valor de $\psi(t)$ ya no cambia (excepto por fluctuaciones), como se muestra en la Fig. 13. La gráfica mostrada en esta figura se generó simulando el sistema de la Fig. 9 en la computadora para una red cuadrada con $N = 10000$ individuos y utilizando un valor de ruido $\eta = 0.05$. Dado que en el límite $t \to \infty$ el valor del parámetro de orden ya no cambia, definimos el *valor estacionario* del

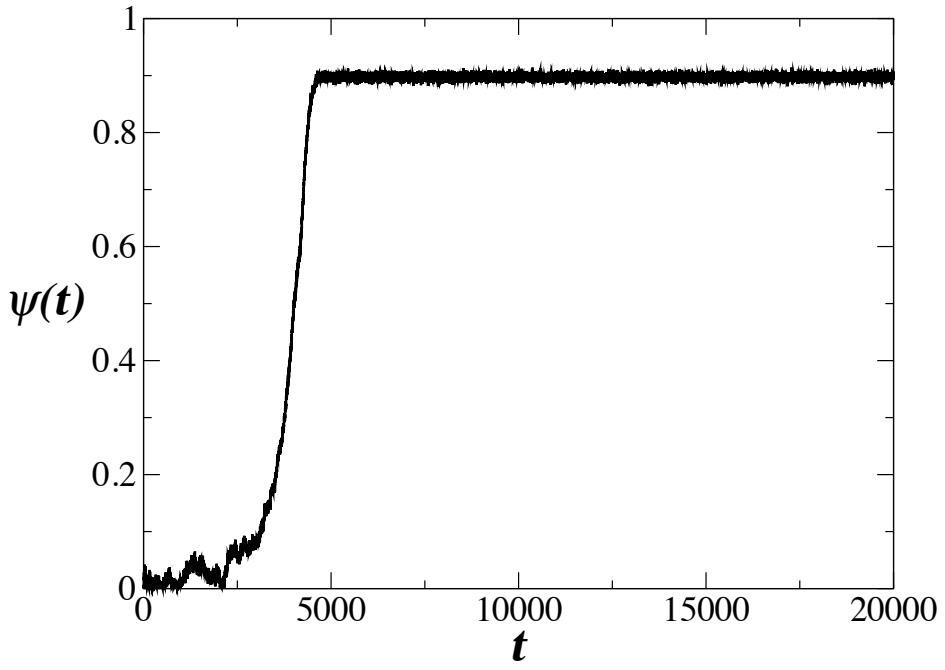


Figura 13: Parámetro de orden $\psi(t)$ como función del tiempo para para la red cuadrada de la Fig. 9 con todos los pesos $w_{n_j} = 1$ y un valor de ruido $\eta = 0.05$. La condición inicial corresponde a un estado en el que la mitad de las opiniones son positivas y la mitad son negativas. Por lo tanto, $\psi(0) \approx 0$. Nótese que después de un tiempo transitorio, aproximadamente en $t = 5000$, el parámetro de orden alcanza un valor constante que ya no cambia en el tiempo (salvo por pequeñas fluctuaciones producidas por el ruido).

parámetro de orden como[8]

$$\psi = \lim_{t\to\infty} \psi(t).$$

Este valor de ψ depende de la cantidad de ruido η presente en el sistema, es decir, de la cantidad de "revoltosos" que existan en la sociedad. La Fig. 14

[8]Estríctamente hablando, la definición correcta es $\psi = \left\langle \lim_{t\to\infty} \psi(t) \right\rangle$, donde los paréntesis angulares $\langle\cdot\rangle$ denotan el promedio sobre todas las posibles condiciones iniciales del sistema. Al tomar este promedio, las fluctuaciones que se observan el la Fig. 13 desaparecen. Otra manera de definir a ψ, que es muy útil para calcularlo numéricamente, es $\psi = \lim_{T\to\infty} \frac{1}{T} \int_0^T \psi(t)\, dt$. La integral lo único que hace es "limpiar" las fluctuaciones producidas por el ruido.

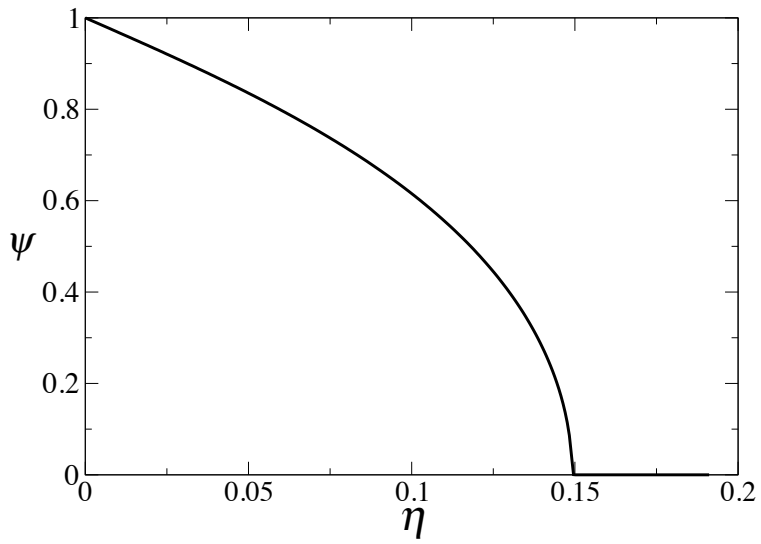


Figura 14: Valor estacionario del parámetro de orden ψ como función del ruido η, para el sistema de la red cuadrada mostrado en la Fig. 9. Obsérvese que al aumentar η, la cantidad de orden disminuye, hasta que para el valor crítico $\eta_c \approx 0.149$ el orden en el sistema se destruye completamente.

muestra el valor estacionario ψ como función del ruido η. Como puede observarse, la cantidad de orden en el sistema es grande para valores pequeños de η. Sin embargo, conforme η aumenta, el orden en el sistema disminuye hasta que, para un valor crítico $\eta_c \approx 0.149$ el orden en el sistema desaparece completamente. Lo que la Fig. 14 está mostrando es una *transición de fase dinámica* de segundo orden (análoga a la mostrada en la Fig. 8). Esta transición de fase confirma las observaciones que hicimos en la sección anterior: Pequeñas cantidades de ruido son necesarias para generar orden en el sistema. Sin este ruido simplemente no se genera una opinión mayoritaria en la sociedad. Pero mucho ruido destruye el orden, dejándonos con una sociedad indecisa.

En este capítulo vimos los conceptos básicos referentes a los procesos dinámicos en redes. Utilizamos una red muy simple, la red cuadrada en la que cada nodo tiene 5 inputs (los 4 primeros vecinos y él mismo). El análisis se puede extender a redes con topologías más realistas, como la red libre de escala o la de mundo pequeño, y utilizando pesos w_{n_j} todos distintos entre sí. De hecho, el modelo de votantes puede resolverse exactamente de forma

analítica, pero la solución es complicada y va más allá de los objetivos de estas notas introductorias. Sin embargo, aún cuando utilicemos diferentes topologías, o factores de peso distintos, los conceptos fundamentales presentados aquí no cambian. Los que estén interesados pueden encontrar un análisis matemático completo del modelo de votantes en las referencias [11, 12].

12. Criticalidad, robustez y evolución en redes genéticas

"Uno se pregunta [...] cómo es posible que los organismos complejos siquiera evolucionen. Parecen tener tantos genes, tantos efectos múltiples o pleitrópicos de cada gen, tantas posibilidades de que hayan mutaciones letales en el desarrollo embrionario, y todo tipo de problemas debido a su largo su desarrollo".

J.T. Bonner. The evolution of complexity
by means of natural selection. (1988).

Uno de los ejemplos más interesantes de la aplicación de las redes complejas lo encontramos en la evolución de los seres vivos, en particular, cuando tratamos de entender qué es lo que está evolucionando. Déjenme ser más explícito. La teoría de la evolución consta de dos partes. Por un lado está la *evidencia* irrefutable, acumulada en el registro fósil por arqueólogos y antropólogos de todo el planeta, de que las especies de organismos vivos han ido cambiando a lo largo de millones de años. Por lo tanto, el hecho de que la evolución ha ocurrido es incuestionable con toda la evidencia que tenemos ahora. Por otro lado, la teoría de la evolución también trata de explicar los *mecanismos* por medio de los cuales las especies han ido evolucionando. Tales mecanismos aun no están completamente comprendidos y todavía existe mucha controversia respecto a cómo se ha dado la evolución de los seres vivos. Un mecanismo muy importante de evolución lo propuso Darwin, y es la famosa *selección natural*, la cual se ha popularizado de manera trivial diciendo que la selección natural consiste en "*la sobrevivencia del más apto*". Esta formulación de lo que es la selección natural está mal por varias razones, pero la más importante es que genera un razonamiento circular que no conduce a ninguna parte. Porque, ¿cómo se define al "más apto"? Pues precisamente como el que sobrevive. Pero entonces "*la sobrevivencia del más apto*" significaría "*la sobrevivencia del que sobrevive*", lo cual es una definición circular sin ningún valor científico. Como veremos en este capítulo,

entender la evolución de los organismos vivos es mucho más complicado que lo que trae consigo la supuesta sobrevivencia del más apto. Lo que tenemos que entender es cuáles son las características que se seleccionan y cómo se van transformando. Enfocaremos nuestra atención en la evolución de las redes genéticas que determinan los fenotípos de los organismos. A primera vista, podría parecer que esto es un trabajo inacabable, en vista de la gran diversidad de formas de vida que existen en nuestro planeta: plantas, insectos, reptiles, mamíferos; y dentro de cada reino hay cientos de miles o millones de especies distintas. Sin embargo, hay dos propiedades que todos estos organismos tienen en común: han evolucionado en entornos cambiantes, lo cual los ha hecho al mismo tiempo *robustos* y *adaptables*. Comenzaremos entonces definiendo estas dos propiedades con más detenimiento.

La "robustez" y la "evolucionabilidad" (la capacidad de evolucionar) son dos propiedades centrales de los organismos vivos. Sin embargo, siempre es problemático utilizar estas palabras juntas en el mismo título, o incluso en el mismo párrafo, a menos que sea para decir que son propiedades (aparentemente) antagónicas. Por un lado, los organismos vivos son robustos porque pueden mantener su forma y funcionamiento ante la presencia de una gran variedad de perturbaciones, que van desde cambios físicos o químicos transitorios en el entorno, hasta mutaciones permanentes en su material genético. Por lo tanto, la idea de robustez se asocia siempre con la invarianza del fenotipo ante perturbaciones internas (mutaciones genéticas) o externas (cambios ambientales). Por otro lado, los organismos vivos también evolucionan, es decir, cambian su fenotipo adquiriendo nuevas formas y/o funciones, como resultado de las mutaciones genéticas y de las presiones selectivas impuestas por los cambios ambientales. Tanto la robustez como la evolucionabilidad se han observado en diferentes niveles de organización biológica, desde la regulación genética hasta el comportamiento dinámico de ecosistemas enteros. Sin embargo, a pesar de la importancia central de estas dos propiedades para el entendimiento del funcionamiento y evolución de los sistemas biológicos, no es claro aún cuáles son los mecanismos estructurales y dinámicos que generan organismos complejos que son robustos, es decir que no cambian ante perturbaciones, pero que al mismo tiempo son evolucionables, es decir, que tienen la capacidad de eventualmente cambiar y adaptarse como resultado de las mismas perturbaciones.

Estamos tan acostumbrados a que los organismos vivos sean al mismo tiempo robustos y evolucionables, que rara vez consideramos que la existencia simultánea de estas dos propiedades en un mismo sistema dinámico

sea un problema. Sin embargo, si lo contextualizamos en el campo de la informática, o de la ingeniería mecánica, por ejemplo, rápidamente veremos que no es sencillo diseñar ni programas computacionales ni máquinas que sean robustas y evolucionables simultáneamente, es decir que mantengan su funcionamiento ante una gran cantidad de "errores" de los usuarios, y que al mismo tiempo tengan la capacidad de adquirir nuevas funciones (además de las ya existentes) que les permitan adaptarse a dichos errores.

Es importante, por lo tanto, entender cómo la robustez y la evolucionabilidad, dos propiedades aparentemente antagónicas, pueden coexistir en los organismos vivos. Notemos que el problema no consiste en describir mecanismos para la existencia de robustez, o de evolucionabilidad, por separado. Es relativamente sencillo hacer máquinas robustas ante fallas, por ejemplo, por medio de la redundancia, que consiste en duplicar o triplicar las partes que se espera que fallen con mayor frecuencia o que son fundamentales para el funcionamiento de la máquina. Así, si queremos que el sistema de frenos ABS en un coche nunca falle, ponemos dos o tres circuitos controladores idénticos, de tal forma que si uno falla, hay otros para reemplazarlo. Análogamente, no es complicado diseñar sistemas que cambien y se adapten a nuevos retos ambientales. Una red neuronal informática entrenada para reconocer imágenes de botellas se puede volver a entrenar para reconocer imágenes de edificios, por ejemplo.

Así, robustez y evolucionabilidad se pueden tener por separado. Lo que es difícil es entender cómo es que ambas propiedades se manifiestan en un mismo sistema simultáneamente. Un sistema diseñado para ser robusto no cambia, y uno que cambia no es robusto. Sin embargo, los organismos vivos son robustos y sin embargo evolucionan. Como veremos, es la criticalidad dinámica de las redes genéticas la que hace posible esta coexistencia. Estudios teóricos llevados a cabo durante años sobre las propiedades dinámicas de modelos de redes genéticas han mostrado que dichas redes pueden operar en tres fases dinámicas distintas, que se han denominado fase ordenada, fase crítica y fase caótica. En la fase ordenada el sistema es muy robusto a las perturbaciones, pero es incapaz de cambiar. Por otro lado, en la fase caótica incluso perturbaciones muy pequeñas alteran el comportamiento de todo el sistema, es cual es por tanto muy inestable ante perturbaciones. La fase crítica es la que separa la fase ordenada de la fase caótica. Es precisamente en la fase crítica donde se encuentra el balance entre robustez y evolucionabilidad necesario para describir el funcionamiento de los organismos vivos. En la fase crítica el sistema es robusto ante perturbaciones, pero no es tan robusto

como para estar "congelado" evolutivamente. Estas observaciones llevaron a Stuart Kauffman a proponer la hipótesis de "la vida al borde del caos" hace casi 40 años. Desde entonces, no había existido evidencia experimental que soportara esta hipótesis. En este trabajo describiremos la importancia de la criticalidad dinámica a nivel genético para la evolución de los organismos y mostraremos evidencia experimental que confirma la existencia de dicha criticalidad.

12.1. Robustez y evolucionabilidad

No existen definiciones precisas para robustez y evolucionabilidad universalmente aceptadas. Estos términos se definen dependiendo del contexto y del nivel de organización bajo consideración. Por lo tanto, antes de comenzar nuestro estudio, es necesario definir qué entendemos por robustez y evolucionabilidad en el contexto de las redes genéticas. A continuación presentamos lo que parece ser el consenso dentro de la comunidad de Biología de Sistemas:

Definición 2 *Robustez es la invarianza de los fenotipos ante la presencia de perturbaciones internas o externas al organismo.*

Definición 3 *Un sistema biológico se dice que es evolucionable si puede adquirir fenotipos nuevos a través del cambio genético (mutaciones), fenotipos que pueden ayudar al organismo a sobrevivir y reproducirse.*

Para que estas definiciones sean útiles desde un punto de vista cuantitativo, es necesario también definir operacionalmente fenotipo y perturbación. En cuanto a la definición de perturbación no hay problema. Es fácil definirla como la tasa de mutación en el ADN, o como la intensidad del ruido en el nivel de expresión genética. Nosotros en este trabajo utilizaremos una perturbación muy particular, que es la duplicación y divergencia genética. Este proceso consiste en que durante la replicación celular frecuentemente se cometen "errores" en la replicación del ADN, uno de los cuales consiste en que algunos genes se copian por duplicado. El genoma final del organismo tiene entonces dos copias idénticas del mismo gen (o del mismo grupo de genes), lo cual provee cierta redundancia, y por lo tanto cierta robustez, al genoma del organismo, ya que si alguna de las dos copias del gen sufre mutaciones, aún está la otra copia para llevar a cabo la función celular correspondiente. De hecho, rápidamente después de la duplicación genética, una de las dos

copias comienza a mutar, lo cual hace que el gen original y el gen duplicado diverjan en funcionalidad.

Susumo Ohno fue uno de los primeros en percatarse de la importancia de la duplicación y divergencia genéticas en la evolución, ya que este mecanismo constituye una fuente importante de material para la innovación genética. Se estima que al menos el 50 % de los genes en células procariotas, y más del 90 % en células eucariotas, son producto de la duplicación y divergencia genéticas. Por lo tanto, la "perturbación" que estudiaremos en este trabajo consistirá precisamente en duplicar y mutar genes en una determinada red genética, y ver qué consecuencias tiene esto en el fenotipo. Esto nos lleva al segundo punto a definir. Es decir, para saber si el fenotipo es robusto o no ante duplicaciones genéticas, es necesario definirlo de tal forma que se puedan cuantificar sus cambios.

Desde hace varias décadas se formuló la hipótesis de que las características fenotípicas de los organismos están determinadas por los atractores dinámicos de la red genética subyacente. El primero en formular esta hipótesis fue Conrad Waddington, al introducir la metáfora de los paisajes epigenéticos. En esta metáfora, la célula madre embrionaria puede considerarse como una pequeña canica que está a punto de rodar cuesta abajo por un terreno con varias colinas y valles (ver la Fig. 15a). Dependiendo de la condición inicial, la canica puede rodar por un valle o por otro, terminando en un sumidero del que no podrá salir. En el contexto de la célula, cada valle representa una ruta de diferenciación, mientras que cada sumidero representa un destino final o estado funcional final de la célula (ver la Fig. 15b). Así, Waddington se imagina al desarrollo embrionario como un proceso dinámico que tiene diversos atractores y diferentes rutas para llegar a ellos.

La metáfora del paisaje epigenético fue formalizada posteriormente por Stuart Kauffman, quien demostró que modelos simples de redes genéticas efectivamente exhiben atractores dinámicos y diferentes rutas para llegar a ellos. Siguiendo a Waddington, Kauffman formuló la hipótesis de que cada atractor dinámico corresponde a un tipo celular o a un destino celular. Por lo tanto, la información del fenotipo celular no está contenida sólo en la secuencia de bases del ADN, sino también en los atractores dinámicos que se generan a partir de la interacción colectiva de todos los genes en el genoma. Por lo tanto, una definición operativa del fenotipo del organismo puede darse en términos del paisaje epigenético, o lo que es equivalente en el formalismo de Kauffman, en términos del paisaje de atractores dinámicos de la red genética. Volveremos a este punto más adelante.

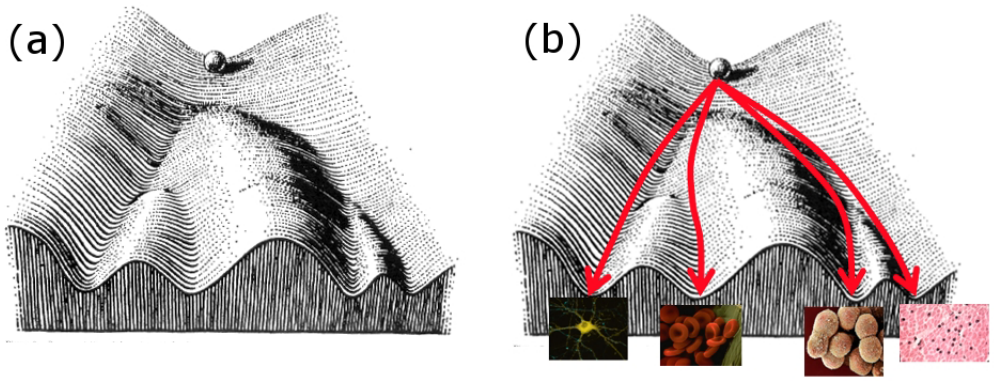


Figura 15: (a) El paisaje epigenético imaginado por Waddington. La célula madre embrionaria corresponde a una canica que está a punto de rodar cuesta abajo por un terreno no uniforme con diversos valles. Dependiendo de la condición inicial (temperatura, presión, acidez, etc.), la canica terminará en alguno de los mínimos estables del terreno. (b) Estos mínimos corresponden a diferentes tipos o destinos celulares, y los valles para llegar a ellos corresponden a las rutas de diferenciación.

12.2. El modelo de Kauffman

Probablemente, el modelo de red genética más sencillo que describe adecuadamente los procesos de regulación genética y diferenciación celular es el que propuso Kauffman en 1969. En este modelo uno está interesado en la evolución temporal del estado de expresión de los genes y no de las concentraciones de sus productos. Por lo tanto, cada gen se representa con una variable binaria σ que puede adquirir los valores $\sigma = 1$ si el gen que representa está expresado, y $\sigma = 0$ si dicho gen no está expresado. El genoma de un organismo se representa entonces por un conjunto de N de estas variables $\{\sigma_1, \sigma_2, \ldots, \sigma_N\}$, cada una correspondiendo a un gen.

Los genes interactúan unos con otros a través de las proteínas reguladoras (factores de transcripción) que generan. Los factores de transcripción se unen de forma específica a las regiones reguladoras de los genes regulados, activando o reprimiendo su expresión. Esta serie de interacciones, mediadas por los factores de transcripción, da lugar a una red dirigida en la que dos genes están "conectados" si la expresión de uno regula, positiva o negativamente, la expresión del otro. La Fig. 16 muestra la estructura topológica de la red de regulación transcripcional de la bacteria Escherichia coli, de acuerdo con

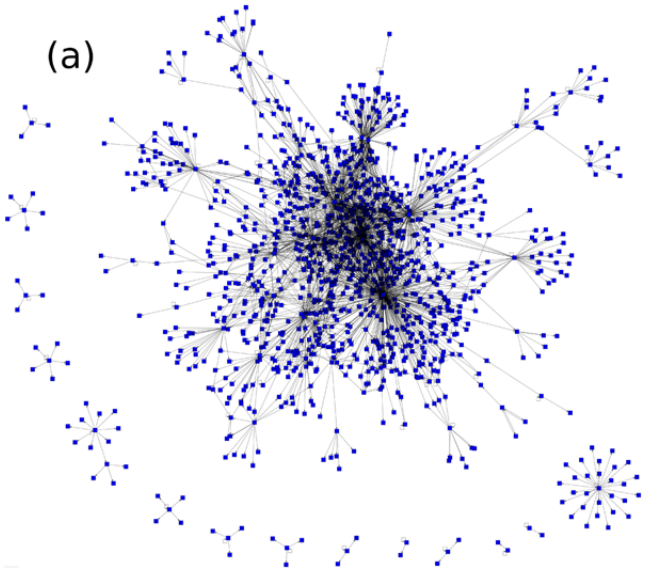


Figura 16: Estructura topológica de la red de regulación transcripcional de *E. coli*. Cada nodo representa un gen, y cada enlace representa una interacción de regulación. Nótese la famosa estructura libre de escala, en la que la gran mayoría de los nodos están poco conectados, mientras que algunos cuantos nodos tienen un gran número de conexiones. Las conexiones son dirigidas porque indican cuáles son los genes reguladores y cuáles son los genes regulados. Sin embargo, la dirección de las flechas no se alcanza a ver en la imagen.

la información más reciente en la base de datos del regulonDB.

En el modelo de Kauffman, a cada gen σ_n se le asocia un conjunto de k_n reguladores, $\{\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}\}$, que son los que determinan el estado de expresión de σ_n. Una vez que se han establecido los reguladores de todos y cada uno de los genes, la dinámica de la red genética queda dada por la ecuación

$$\sigma_n(t+\tau) = F_n\left(\sigma_{n_1}(t), \sigma_{n_2}(t) \ldots, \sigma_{n_{k_n}}(t)\right) \tag{27}$$

Esta ecuación nos dice que el estado de expresión de un gen σ_n al tiempo $t+\tau$ está determinado únicamente por el estado de expresión de sus reguladores $\{\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}\}$ evaluado en un tiempo anterior t. El parámetro τ es una medida del tiempo típico de respuesta de la expresión genética de los genes. La función booleana F_n determina la interacción entre σ_n y sus reguladores, y se construye tomando en cuenta la naturaleza activadora o represora de los reguladores de σ_n. Al iterar la Eq. (27) en el tiempo cada

gen, siguiendo en todo momento su propia función booleana, pasa por una serie de estados prendido/apagado hasta que toda la red entra en un estado de expresión periódico (ver la Fig. 17). Algunos genes alcanzan una expresión constante que ya no cambia en el tiempo, mientras que otros genes se siguen prendiendo y apagando pero de forma periódica. Este estado de expresión periódico es el atractor dinámico de la red al que nos referimos antes. Generalmente, una misma red genética tiene varios atractores dinámicos. Cada atractor dinámico queda identificado de forma única por el conjunto de genes que están prendidos en él. En otras palabras, en atractores distintos se expresan conjuntos de genes diferentes. Pero ésta es precisamente la característica que identifica a los diferentes tipos celulares en un organismo, es decir, *conjuntos distintos de genes se expresan en tipos celulares distintos.* Por tal motivo, Kauffman formuló la hipótesis de que *los atractores dinámicos de la red genética corresponden a los diferentes tipos o destinos celulares observados en el organismo.*

Esta hipótesis ha sido confirmada recientemente con estudios numéricos y experimentales llevados a cabo por diversos grupos de investigación, entre los que destacan el de la Dra. Elena lvarez-Buylla del Instituto de Ecología de la UNAM, quienes encontraron que los atractores de la red de desarrollo floral de la planta Arabidopsis thaliana corresponden a los diferentes patrones de expresión genética observados experimentalmente en los distintos órganos de la planta. Estudios similares llevados a cabo por el grupo de la Dra. Reka Albert de la Universidad de Pensilvania, el grupo del Dr. Sui Huang del Harvard Children's Hospital, y el grupo del Dr. Chao Tang de la Universidad de California en San Francisco, muestran también que los atractores de la red genética de diferentes organismos (*Drosophila melanogaster*, células de tejido endotelial, y *Saccharomyces cerevisiae*) corresponden a los patrones de expresión observados en diferentes tipos o destinos celulares. Por lo tanto, estos trabajos muestran de forma contundente que características fenotípicas importantísimas del organismo están codificadas en los atractores dinámicos de la red genética subyacente.

12.3. El paisaje de atractores

El paisaje de atractores en el modelo de Kauffman es equivalente al espacio fase de un sistema dinámico. Es una manera de representar gráficamente las diferentes trayectorias que sigue el sistema al moverse de acuerdo con la ecuación que dicta su dinámica, en nuestro caso la Eq. (27). Para ilustrar

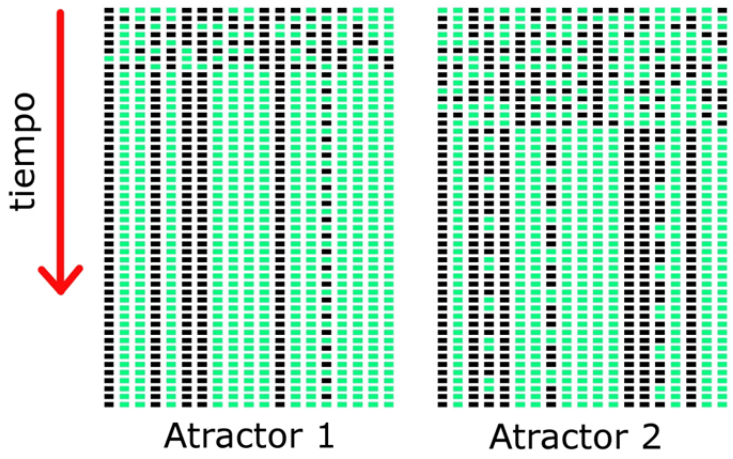


Figura 17: Representación visual de la dinámica en el modelo de Kauffman. Cada cuadrito representa un gen, en color claro si el gen está expresado y en negro si no está expresado. Cada fila horizontal representa el estado de todo el genoma en un instante de tiempo. La primera fila es la condición inicial con la que se comienza la dinámica, la segunda fila es el estado del genoma después de una iteración de la Eq. (27), la segunda fila después de dos iteraciones, y así sucesivamente. Puede observarse que después de un tiempo transitorio el genoma entra en un patrón de expresión periódico, que es el atractor dinámico. Los dos paneles mostrados corresponden a diferentes condiciones iniciales que terminan en diferentes atractores.

la construcción del paisaje de atractores en el modelo de Kauffman, supongamos que tenemos una red genética de sólo 4 genes. Existen $\Omega = 2^4 = 16$ estados dinámicos posibles en los que la red se puede encontrar, desde el estado 0000 en el que todos los genes están apagados, hasta el estado 1111 en el que todos los genes están prendidos, pasando por todos los estados intermedios. Todos estos estados posibles de expresión en los que se puede encontrar la red en algún momento se muestran en la Fig. 18a. A lo largo del tiempo, el estado de expresión de cada gen cambia de acuerdo a la Eq. (27), pero por el momento no consideraremos los cambios que sufre cada gen, sino los cambios que sufre la red completa, es decir, consideraremos el efecto neto de todas las funciones booleanas sobre el estado de toda la red. Supongamos como indica la Fig. 18b que, al iniciar la dinámica con la condición inicial 0000, después de un paso de tiempo la red cambia al estado 0101, el cual a su vez cambia al estado 1100 en el siguiente paso de tiempo, el cual se convierte en 0110 que después vuelve a cambiar a 0101. A partir de este momento se

vuelve a repetir el ciclo ya que el estado 0101 ya se había encontrado antes. En este ejemplo, la red sigue la trayectoria en el espacio de estados

$$0000 \to 0101 \to 0110 \to 1100 \to 0101 \to 0110 \to 1100 \to \cdots$$

El atractor consiste de los tres estados 0101, 0110 y 1100. Si ahora comenzáramos la dinámica con la condición inicial 1110, se tendría la trayectoria

$$1110 \to 1011 \to 0010 \to 0011 \to 0011 \to 0011 \to 0011 \to \cdots$$

En este otro caso el atractor consiste de un sólo estado, 0011. La Fig. 18c muestra el diagrama al que se llegaría si conectamos con flechas todos los posibles estados dinámicos de la red conforme se van sucediendo uno al otro en el tiempo hasta llegar al atractor. Todos los estados que eventualmente convergen al mismo atractor conforman la cuenca de atracción de ese atractor. Como se ve en la Fig. 18c, la misma red puede tener más de un atractor. La Fig. 18d muestra un ejemplo similar pero para una red con 10 genes, la cual tiene entonces $\Omega = 2^{10} = 1024$ estados (una red con N genes tendrá $\Omega = 2^N$ estados posibles).

Nótese como la dinámica generada por la Eq. (27) divide al espacio de estados en los diferentes atractores y sus respectivas cuencas de atracción. Las estructuras que resultan de esta división, como las mostradas en la Fig. 18c y 18d, es lo que se conoce como el paisaje de atractores. Desde el punto de vista biológico, los estados que conforman a los atractores son patrones de expresión genéticos que contienen la información del tipo celular (riñón, hígado, pulmón, neurona, etc.), mientras que las "ramas" que convergen a esos atractores son las diferentes rutas de diferenciación. Es hasta cierto punto sorprendente que un modelo tan sencillo como el propuesto por Kauffman pueda reproducir acertadamente dos de los mecanismos fundamentales de los sistemas biológicos, que es la regulación genética y la diferenciación celular.

12.4. Criticalidad en el modelo de Kauffman

Las redes de Kauffman pueden operar en tres regimenes dinámicos distintos: la fase ordenada, la fase crítica y la fase caótica. La fase crítica es donde se da la transición entre el orden y el caos. Antes de describir en qué consisten

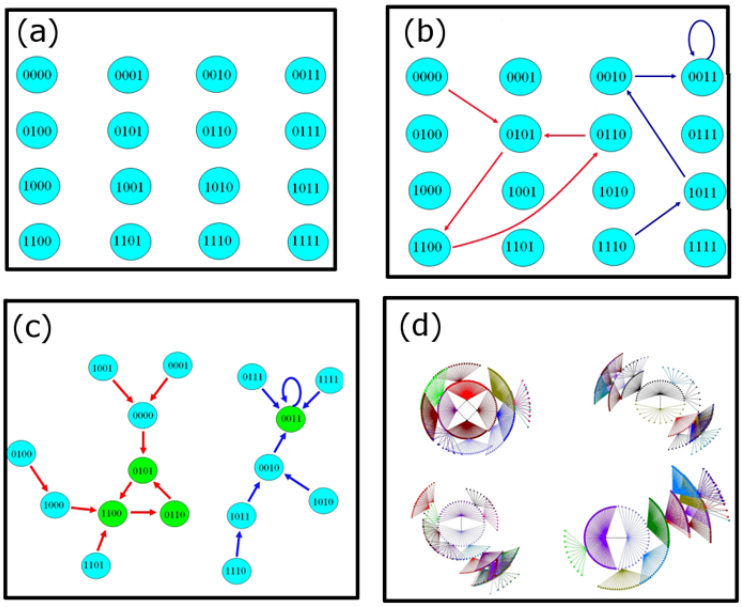


Figura 18: Representación gráfica del paisaje de atractores. (a) Una red con sólo cuatro genes puede estar en cualquiera de los 16 estados posibles listados aquí. (b) La dinámica generada por la Eq. (27)) conecta los estados en el tiempo, de tal manera que dos estados están conectados si uno es sucesor del otro. (c) Al conectar todos los posibles estados con sus sucesores, observamos que hay sólo dos atractores: uno que consiste de los tres estados $\{0101, 1100, 0110\}$, y otro que consiste de un solo estado $\{0011\}$. Los atractores se pintan en color verde y las cuencas de atracción en color azul. (d) Un ejemplo similar pero para una red con 10 genes y 1024 estados. Es esta caso hubieron cuatro atractores. Las estructuras en forma de abanico reflejan el hecho de que varios estados pueden tener el mismo sucesor.

estas tres fases, es importante mencionar que existe evidencia experimental y teórica de que muchos sistemas complejos dinámicos, además de las redes genéticas, operan en la fase crítica. Ejemplos típicos son avalanchas de nieve o de arena, precipitación atmosférica, mercados financieros, terremotos, medios granulares, y la actividad neuronal en el cerebro, por mencionar sólo algunos. El interés en los sistemas críticos, es decir los que operan al borde del orden y del caos, radica en que tales sistemas presentan propiedades notables que serían muy difíciles de explicar en ausencia de la criticalidad. Por ejemplo, tales sistemas pueden integrar, procesar y transferir información más rápidamente y de forma más eficiente que sistemas no críticos. O

pueden detectar y responder a estímulos cuyas intensidades varían en varios órdenes de magnitud[9]. Estas propiedades de los sistemas críticos son principalmente una consecuencia de las correlaciones de largo alcance que emergen en la fase crítica, generando comportamientos colectivos del sistema como un todo. Así, la criticalidad le confiere al sistema la propiedad de responder y adaptarse como un todo a las fluctuaciones o perturbaciones que ocurren en entornos rápidamente cambiantes.

En el modelo de Kauffman, la fase dinámica en la que opera la red se caracteriza por el comportamiento temporal de la distancia Hamming $x(t)$ entre dos trayectorias diferentes en el espacio de estados. Supongamos que inicializamos la red en alguna condición en la que algunos genes están prendidos y otros apagados. Llamaremos a esta condición la condición silvestre. Bajo de dinámica dada por la Eq. (27), la red seguirá una trayectoria como la indicada en el panel izquierdo de la Fig. 19, a la que llamaremos trayectoria silvestre (el primer renglón de dicha trayectoria es la condición inicial). Ahora supongamos que inicializamos la red en una condición ligeramente diferente a la anterior, en la que algunos genes tienen un estado de expresión diferente respecto a la condición silvestre. Esto se puede lograr aumentando o disminuyendo artificialmente el nivel de expresión de algunos genes. De cualquier forma, los genes cuya expresión ha sido alterada respecto a la condición silvestre se han marcado con rojo en el panel derecho de la Fig. 19 (los dos genes rojos en el primer renglón). Esta nueva condición "perturbada" generará una trayectoria en la que la perturbación inicial puede cambiar la expresión de otros genes a lo largo del tiempo, propagándose a otras partes de la red. Es decir, la perturbación inicial generará una avalancha de perturbaciones que puede o no propagarse a otros elementos de la red. En el ejemplo mostrado en la Fig. 19, los genes que han sido perturbados respecto a la trayectoria silvestre se han marcado en rojo.

Definición 4 *La distancia Hamming $x(t)$ es la fracción de genes perturbados al tiempo t en toda la red, es decir, es la fracción de genes en el genoma que han cambiado su nivel de expresión (respecto a la trayectoria silvestre) debido a la perturbación inicial.*

[9]Por ejemplo, el cerebro puede detectar y responder a sonidos cuyas intensidades varían desde unos 5 decibeles (ruido de una hoja movida por el viento) hasta 130 decibeles (bocinas en una tocada callejera de rock). Esto representa una diferencia en intensidades de 12 órdenes de magnitud. Lo mismo pasa con la vista.

De ahora en adelante nos referiremos a $x(t)$ como *el tamaño (fraccional) de la avalancha de perturbaciones*. La evolución temporal de $x(t)$ está dada por un mapeo dinámico $x(t+\tau) = M(x(t))$ el cual relaciona el tamaño de la avalancha de perturbaciones en dos pasos de tiempo consecutivos. Dada una perturbación inicial $x(0)$ al tiempo $t = 0$, iteraciones sucesivas de este mapeo eventualmente convergerán a un valor final $x_\infty = \lim_{t\to\infty} x(t)$, que es el valor final del tamaño de la avalancha. Es precisamente este valor final x_∞ el que caracteriza la fase dinámica en la que opera la red. Si $x_\infty = 0$ entonces la avalancha de perturbaciones eventualmente desaparece, y por lo tanto la expresión del genoma regresa a su estado silvestre no perturbado. Esta es la fase ordenada en la cual las perturbaciones transitorias no tienen ningún efecto en el estado de expresión estable de la red. Por el contrario, si $x_\infty > 0$ entonces la perturbación inicial de unos cuantos genes se propaga a toda la red, alterando eventualmente la expresión de una fracción considerable de genes. Esta es la fase caótica en la que la expresión de todo el genoma cambia erráticamente ante perturbaciones. Finalmente, en la fase crítica, el comportamiento típico de la avalancha de perturbaciones es que no crece ni decrece, sino que se queda más o menos del mismo tamaño. Así, una pequeña perturbación inicial permanecerá pequeña a lo largo del tiempo (o desaparecerá), pero no se propagará a todo el sistema. En esta fase la dinámica de la red no es extremadamente sensible a perturbaciones, pero tampoco está congelada. El comportamiento de la avalancha de perturbaciones se ilustra gráficamente en la Fig. 20.

En 1986 Y. Pomeau y B. Derrida formularon una teoría de campo medio que demuestra la existencia de la transición de fase entre orden y caos en el modelo de Kauffman. En esta teoría se asume que cada gene tiene en promedio K reguladores, seleccionados aleatoriamente con probabilidad uniforma de entre todos los genes que conforman la red (esta elección da lugar a la topología homogénea aleatoria). Además, se asume que todos los genes son estadísticamente independientes y estadísticamente equivalentes. Con estas suposiciones, se puede demostrar que el mapeo $M(x)$ que determina la dinámica temporal del tamaño de la avalancha de perturbaciones tiene las siguientes propiedades (la demostración matemática de estas propiedades puede encontrarse en la referencia [[13]]):

- $M(x)$ es una función continua monótonamente creciente en el intervalo $x \in [0, 1]$.

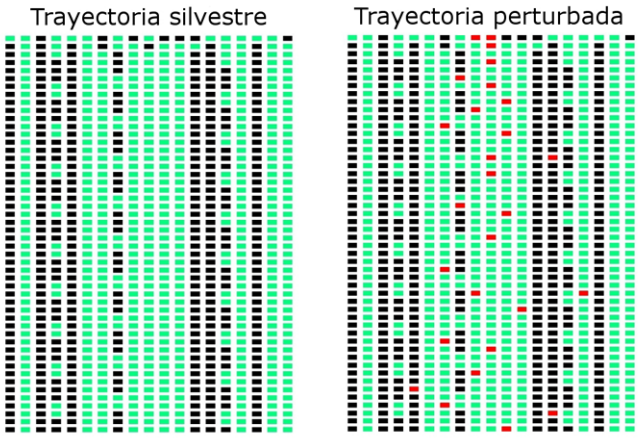


Figura 19: La distancia Hamming es el tamaño fraccional de la avalancha de perturbaciones producida por una perturbación inicial. El panel izquierdo muestra la "trayectoria silvestre" respecto de la cual se medirán las perturbaciones. En el panel derecho se muestra la trayectoria perturbada. Los genes en rojo son aquellos que han cambiado su nivel de expresión, respecto de la trayectoria silvestre, debido a la perturbación inicial (que consiste en los dos cuadritos rojos en el primer renglón).

- $M(0) = 0$, es decir, si no hay perturbación al tiempo $t = 0$ entonces no habrá perturbación a ningún tiempo posterior.
- $0 < M(1) < 1$, es decir, el mapeo siempre se mantiene acotado en el intervalo $x \in [0, 1]$.

La forma general de este mapeo se muestra en la Fig. 21 para tres casos ilustrativos. En el primer caso (primer panel de la Fig. 21) el mapeo $M(x)$ va completamente por debajo de la identidad. Por lo tanto, dada una perturbación inicial x_0, iteraciones sucesivas de este mapeo convergerán rápidamente hacia $x_\infty = 0$ y la perturbación desaparecerá. Esta es la fase ordenada. En el segundo caso (segundo panel de la Fig. 21), el mapeo $M(x)$ comienza en el origen por arriba de la diagonal y la cruza en algún punto $x_\infty < 1$. Por lo tanto, iniciando con una pequeña perturbación inicial x_0, iteraciones sucesivas de este mapeo amplifican la perturbación hasta que alcanza un tamaño final x_∞ que es más grande que x_0. Esta es la fase caótica. El tercer caso (tercer panel de la Fig. 21) es cuando el mapeo $M(x)$ es tangente a la diagonal en el origen. En este último caso una pequeña perturbación inicial x_0 comienza

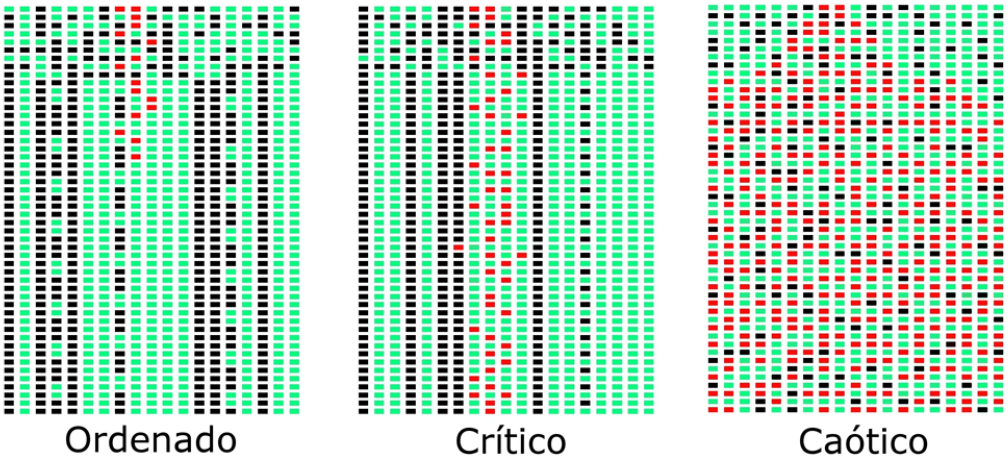


Figura 20: Las tres fases se caracterizan por la propagación de las perturbaciones a través del sistema. En la fase ordenada, una pequeña perturbación inicial (cuadritos rojos en el primer renglón) rápidamente desaparece en el tiempo. En la fase crítica la perturbación inicial no desaparece, pero tampoco crece, se mantiene confinada. En la fase caótica la perturbación se propaga a todo el sistema.

disminuyendo, pero debido a que el mapeo es tangente a la diagonal en el origen, tardará muchísimo tiempo en desaparecer. Para todos los tiempos finitos parece que la perturbación inicial x_0 no crece ni decrece en el tiempo y se tendrá entonces que $x_\infty \approx x_0 \approx 0$. Esta es precisamente la característica de la fase crítica.

Las observaciones anteriores muestran que lo que determina la fase dinámica en la cual opera la red es la pendiente del mapeo $M(x)$ en el origen, es decir, en $x = 0$. Debido a que este es un parámetro importante, se le ha dado un nombre: *la susceptibilidad de la red*, y se le denota con el símbolo S. Si $S < 1$ la red está en la fase ordenada (el mapeo comienza por debajo de la identidad); si $S > 1$ la red está en la fase caótica (el mapeo comienza por arriba de la identidad). La fase crítica se obtiene cuando $S = 1$ (el mapeo comienza tangente a la identidad). La teoría de campo medio predice que la susceptibilidad de la red está determinada por la ecuación

$$S = 2p(1-p)K \tag{28}$$

donde K es el número promedio de reguladores por gen en toda la red, y

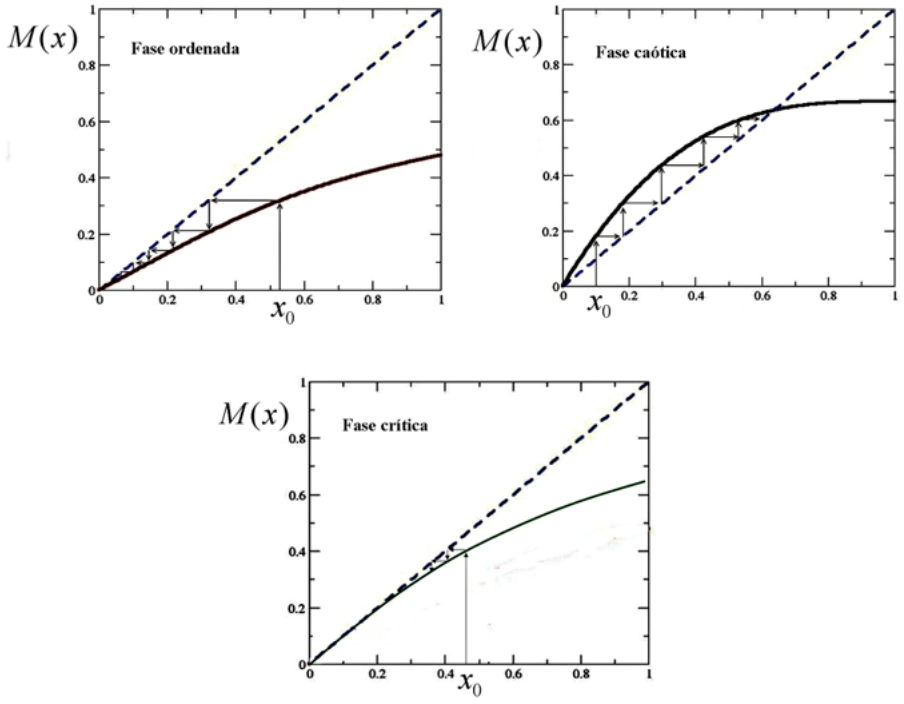


Figura 21: Formas típicas del mapeo $M(x)$ en las tres fases dinámicas. En cada caso las flechas indican la evolución temporal de la perturbación al iterar el mapeo. En la fase ordenada una perturbación inicial x_0 desaparece rápidamente. En la fase caótica la pequeña perturbación inicial x_0 se amplifica. En la fase crítica la perturbación inicial x_0 comienza a disminuir, pero rápidamente queda "atrapada" entre la recta identidad y el mapeo, pareciendo como si no disminuyera ni aumentara.

p es la fracción promedio de frases de regulación activadoras. El parámetro K está relacionado con la topología de la red, mientras que el parámetro p tiene que ver con las funciones booleanas que regulan la actividad de los genes. Así, tanto la topología de la red genética como la naturaleza de las interacciones entre los genes determinan si la red opera en la fase ordenada, en la crítica o en la caótica.

12.5. Criticalidad en redes genéticas de organismos reales

La teoría de campo medio desarrollada por Pomeau y Derrida hace suposiciones que evidentemente no son ciertas para los organismos reales. En particular, las suposiciones de que los genes son estadísticamente independientes y estadísticamente equivalentes no se cumplen. Como muestra la Fig. 16, hay genes que están altamente conectados y genes que están muy poco conectados. Por lo tanto, los genes no son equivalentes. Además, los genes altamente conectados, también llamados reguladores globales, regulan la expresión de cientos de otros genes, lo cual introduce correlaciones en la expresión genética de gran parte de la red. La existencia de los reguladores globales hace que la suposición de campo medio sobre la independencia estadística de los genes tampoco sea cierta. Por lo tanto, es importante determinar, en primer lugar, si la transición de fase entre orden y caos que describimos anteriormente, y que predice la teoría de campo medio, sigue siendo válida para redes cuya topología es como la observada en las redes genéticas de organismos reales. La Fig. 22 muestra el valor final del tamaño de la avalancha de perturbaciones, x_∞ , como función de la susceptibilidad S de la red. La línea negra sólida corresponde a lo predicho por la teoría de campo medio, mientras que la línea roja quebrada es el resultado que se obtiene al simular en la computadora la dinámica de Kauffman en redes con topologías "realistas", es decir, construidas con las mismas propiedades topológicas que las observadas en redes de organismos reales (como la mostrada en la Fig. 16). Puede observarse en la Fig. 22 que la transición de fase predicha por la teoría de campo medio es idéntica a la obtenida para redes con topologías realistas, aún cuando para tales redes las hipótesis del campo medio no se cumplen. Este resultado es bastante inesperado y muestra un ejemplo en el cual la teoría de campo medio conduce a resultados correctos incluso con suposiciones francamente equivocadas.

Pare determinar si las redes genéticas de organismos reales operan en la fase crítica, lo que hicimos fue calcular el mapeo $M(x)$ utilizando los datos experimentales que se tienen hasta el momento de las redes genéticas de 5 organismos: la planta *Arabidopsis thaliana*, la mosca *Drosophila melanogaster*, la bacteria *Escherichia coli*, la bacteria *Bacillus subtilis* y el hongo *Saccharomyces cerevisiae*. En el caso de la planta utilizamos la red de desarrollo floral que determina el surgimiento de los diferentes órganos de la planta (pétalos, estambres, carpelos, etc.). Para la mosca utilizamos la red que determina la segmentación polar del embrión en las diferentes partes

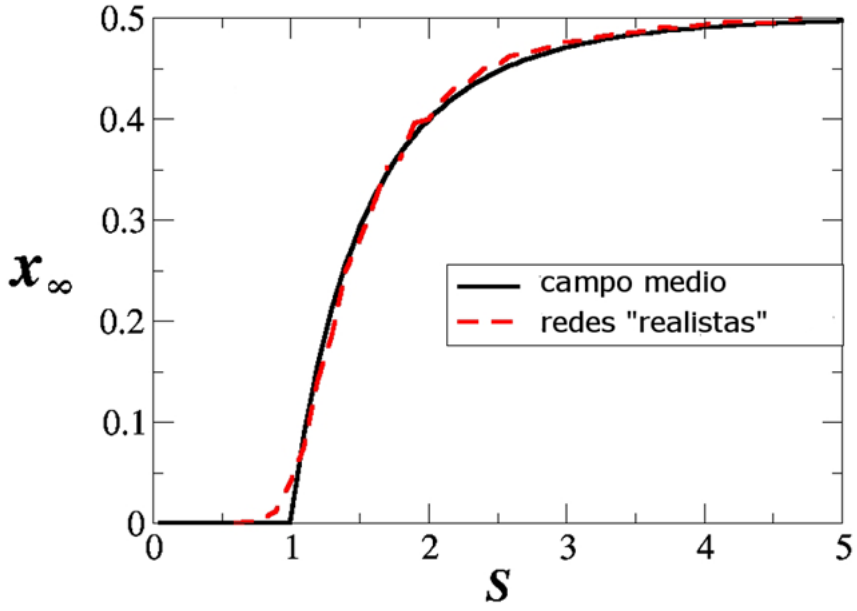


Figura 22: Transición de fase predicha por la teoría de campo medio (curva negra continua) y la obtenida numéricamente para redes con topologías "realistas" (curva roja quebrada). Nótese como la transición de fase es idéntica en ambos casos.

que eventualmente darán lugar a distintos órganos. Estas dos redes están completamente documentadas y basadas en datos experimentales, tanto la topología como las funciones booleanas reguladoras. No son redes grandes[10], pero son módulos genéticos representativos dentro del genoma encargados de llevar a cabo funciones bien caracterizadas.

Por otro lado, las redes que utilizamos para los organismos unicelulares *E. coli*, *B. subtilis* y *S. cerevisiae* son redes mucho más grandes[11] que intentan describir todas las interacciones transcripcionales en los genomas de esos organismos. Aunque dichas redes todavía no están completas, las partes que ya se tienen en las bases de datos son lo suficientemente grandes como para ser estadísticamente representativas de la estructura de toda la red. El problema es que para estas redes sólo se conoce la topología, pe-

[10]La red del desarrollo floral de *A. thaliana* tiene 15 genes, y la red de segmentación polar de *D. melanogaster* tiene 60 genes.

[11]Actualmente, las bases de datos cuentan con 1400, 900 y 3500 genes, respectivamente.

ro no se conocen las funciones booleanas. Por lo tanto, en ausencia de este conocimiento lo que hicimos para calcular el mapeo $M(x)$ fue implementar funciones booleanas aleatorias. Sin embargo, las funciones aleatorias que utilizamos fueron construidas utilizando como modelo la expresión genética a gran escala observada en experimentos de microarreglo. Dichos experimentos muestran imágenes "instantáneas" del estado de expresión genética de todos los genes en un organismo (ver la Fig. 23). Recabamos todos los experimentos de microarreglo que pudimos encontrar en las bases de datos públicas (fueron más de 400 en total), y de ellos inferimos el valor del parámetro p que aparece en la Eq. (28), es decir, el valor de la fracción de frases de regulación activadoras en todo el genoma. Con este valor de p construimos aleatoriamente las funciones booleanas que después implementamos en las redes de *E. coli*, *B. subtilis* y *S. cerevisiae*. Por lo tanto, aún cuando las funciones booleanas que utilizamos eran aleatorias con un sesgo dado por p, las construimos de la forma más compatible posible con los datos experimentales existentes.

El mapeo $M(x)$ obtenido para las redes genéticas de los 5 organismos mencionados anteriormente se muestra en la Fig. 24. Como puede observarse, en todos los casos el mapeo es consistente con dinámicas críticas. Aún cuando la sensitividad S de la red no es exactamente igual al valor crítico $S = 1$ en ningún caso, está bastante cerca de dicho valor crítico. Esto es particularmente cierto para las redes de *E. coli* y *S. cerevisiae*, que son las más completas que tenemos hasta el momento. Los datos reportados en la Fig. 24 constituyen la evidencia más sólida de criticalidad en la dinámica genética de organismos reales encontrada hasta la fecha.

12.6. Criticalidad, robustez y evolución

La evidencia que hemos presentado hasta el momento sobre la existencia de criticalidad dinámica en redes genéticas está basada principalmente en el estudio de la respuesta de la red ante perturbaciones transitorias en el estado de expresión de los genes. Es decir, la fase dinámica en la cual opera la red determina si ésta será capaz o no de recuperarse de la sub-expresión o de la sobre-expresión de unos cuantos genes inducida por fluctuaciones en el medio ambiente. Sin embargo, existe una manifestación mucho más fundamental de la criticalidad, y tiene que ver con la transformación del paisaje de atractores ante mutaciones permanentes de la red.

Como mencionamos en la introducción, el principal mecanismo de evolución y crecimiento genómico es la duplicación genética seguida por diver-

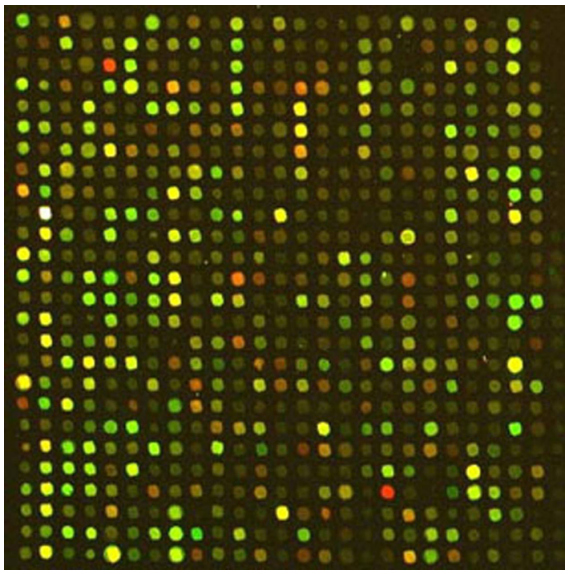


Figura 23: Con la tecnología existente se puede observar, en tiempo real, el estado de expresión de todos los genes en el genoma. En esta imagen se muestra un experimento de microarreglo en el que cada punto brillante indica el estado de expresión de un gen distinto en *E. coli*. Entre más brillante sea el punto, más activo está el gen correspondiente. Utilizando estadística bayesiana y la topología conocida de la red genética, se puede inferir, a partir de estos experimentos, la fracción de frases booleanas activadoras en todo el genoma.

gencia. Sin embargo, el paisaje de atractores depende de forma no trivial del número de genes que hay en la red y de todas las interacciones entre ellos. Por lo tanto, duplicar un gen y modificarlo puede alterar fuertemente el paisaje de atractores. No queremos que esto ocurra ya que el paisaje de atractores contiene información de las características fenotípicas más relevantes del organismo. Por lo tanto, es importante determinar hasta qué punto el proceso de duplicación y divergencia genética modifica el paisaje de atractores. Esto nos lleva a la siguiente definición operacional de la robustez y evolucionabilidad genéticas:

Definición 5 *Una red de Kauffman con N_a atractores es robusta si todos sus M_a atractores se conservan después de duplicar y mutar un sólo gen. Adicionalmente, diremos que la red es evolucionable si, como consecuencia de la duplicación y divergencia de un gen, nuevos atractores emergen.*

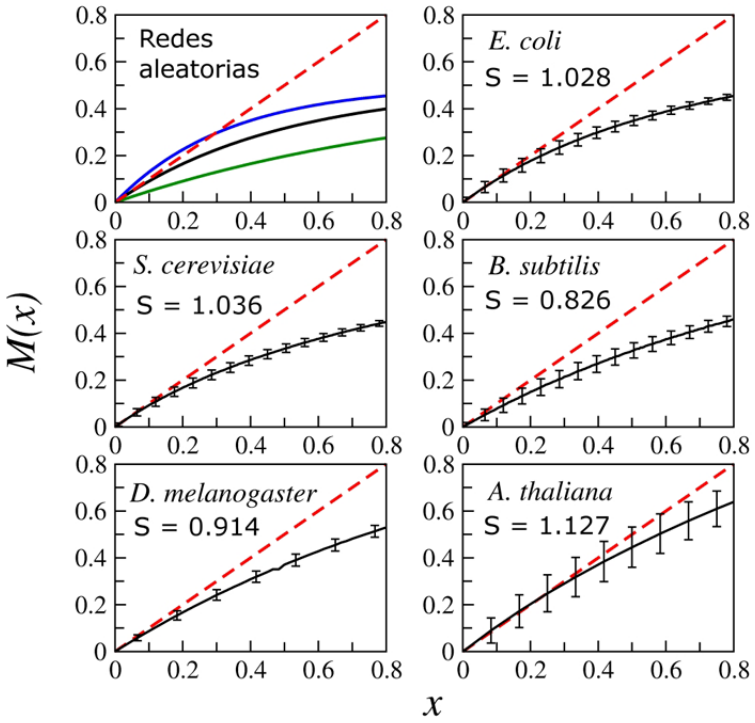


Figura 24: El mapeo $M(x)$ muestra criticalidad dinámica en las redes genéticas de 5 organismos modelo ampliamente estudiados en biología, los cuales representan 4 reinos distintos de la vida. El valor de la susceptibilidad S (la pendiente del mapeo en el origen) se indica. Para facilitar la comparación, el panel superior izquierdo muestra el mapeo $M(x)$ para distintas redes aleatorias operando en las tres fases dinámicas.

Vale la pena enfatizar que el paisaje de atractores depende de las interacciones específicas que han entre todos los genes de la red, es decir, que el paisaje de atractores no es una propiedad particular de un solo gen, o de un grupo reducido de genes, sino que emerge del comportamiento colectivo de toda la red. Por lo tanto, la robustez y evolucionabilidad definidas anteriormente son también propiedades emergentes colectivas asociadas a toda la red, o como lo diría A. Wagner, son propiedades *distribuidas*.

La duplicación genética seguida de divergencia se puede implementar fácilmente en el modelo de Kauffman. Comenzamos con una red que tiene N genes, cada uno de los cuales tiene asociado un conjunto de reguladores y una función booleana. Esta red, a la cual llamaremos *la red original*, tiene un determinado paisaje de atractores que determinan las características

fenotípicas del organismo. De la red original seleccionamos un gen al azar y lo duplicamos, por lo tanto, el gen duplicado tendrá los mismos reguladores (inputs), los mismos genes regulados (outputs) y la misma función booleanas que el gen del que proviene, tal y como se ilustra en la Fig. 25. Esto producirá una red con $N + 1$ genes, siendo el último gen una copia idéntica de alguno de los genes originales. Después simulamos el proceso de divergencia genética mutando al gen duplicado, es decir, cambiando permanentemente alguno de sus reguladores, alguno de sus regulados, o su función booleanas (o todo al mismo tiempo, lo cual equivaldría a añadir un gen completamente nuevo a la red). Nos referiremos a la red que resulta después de este proceso de duplicación y divergencia como *la red mutada*. No hay razón *a priori* para suponer que los atractores de la red original serán iguales a los de la red mutada. Se sabe, por ejemplo, que el número promedio de atractores aumenta con el tamaño de la red, así que esperamos que la red mutada tenga, en promedio, más atractores que la red original[12].

Lo que nos interesa es comparar el paisaje de atractores de la red original con el paisaje de atractores de la red mutada, y cuantificar los cambios. En particular, queremos saber cuántos atractores de la red original se conservan, y cuántos cambian. Para esto, generamos en la computadora poblaciones de decenas de miles de redes distintas, cada red con su propio paisaje de atractores. En cada red implementamos el proceso descrito en la Fig. 25 y comparamos los atractores de la red original con los atractores de la red mutada. La Fig. 26 muestra la fracción $P(q)$ de redes en la población en las cuales un porcentaje q de los atractores originales se conservaron en la correspondiente red mutada. Por ejemplo, si $P(30) = 0{,}4$, esto quiere decir que en dos quintas partes de la población de redes, el 30 % de los atractores originales se conservaron después de la duplicación y divergencia, mientras que el 70 % de los atractores originales cambiaron. La Fig. 26 muestra resultados para poblaciones de redes operando en las tres fases dinámicas. Puede observarse en esta figura que para las redes ordenadas y para las críticas sobresalen dos picos en la gráfica de $P(q)$, uno en $q = 0\,\%$ y el otro en $q = 100\,\%$, y casi no hay nada entre estos dos extremos. Esto quiere decir que para tales redes, después de la duplicación y divergencia de un solo gen, ningún atractor se conservó, lo cual corresponde al pico en $q = 0\,\%$, o bien todos los atractores originales se conservaron, lo cual corresponde al pico en $q = 100\,\%$. Casi no hay casos

[12]Esto es consistente con el hecho de que organismos más complejos (con más características fenotípicas) tienen en general genomas más grandes.

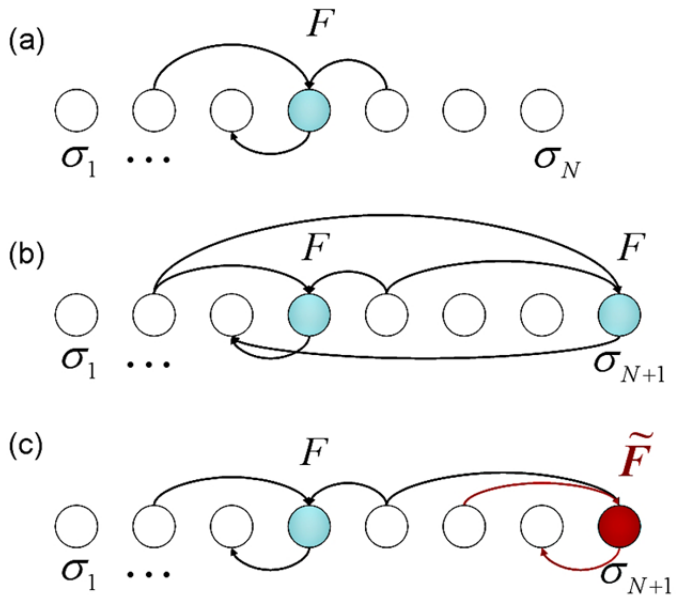


Figura 25: Duplicación y divergencia genética en el modelo de Kauffman. (a) La red original con N genes. Se selecciona un gen particular, el cual se muestra con sus inputs, outputs y función booleana F. (b) El gen seleccionado se duplica, lo cual da lugar a una red con $N+1$ genes. El gen duplicado es idéntico al gen padre en el sentido de que tiene los mismos inputs, outputs y la misma función booleana. (c) Algunas características del gen duplicado se cambian, como sus inputs, sus outputs o su función booleanas, lo cual hace que el gen duplicado ya no sea igual al gen padre. La red mutada no sólo tiene un gen adicional, sino que dicho gen puede ser funcionalmente diferente a todos los genes de la red original.

intermedios. O se conserva todo, o cambia todo. Sin embargo, notemos de la Fig. 26 que la probabilidad de que se conserven todos los atractores es más grande que la probabilidad de que ninguno se conserve. Por lo tanto, para redes operando en las fases ordenada o crítica, lo más probable es que, después de un evento de duplicación y divergencia se mantengan las características fenotípicas del organismo codificadas en los atractores dinámicos de la red. Esta es la robustez del paisaje epigenético que mencionamos anteriormente. Dicha robustez no está presente en las redes operando en la fase caótica, como puede observarse en la Fig. 26, ya que en dicha fase el pico en $q = 100\,\%$ desaparece, mientras que el pico en $q = 0\,\%$ no sólo no desaparece, sino que

es el más prominente. Por lo tanto, en la fase caótica no hay robustez ya que el paisaje de atractores nunca se conserva.

¿Y que pasa con la evolucionabilidad? Como habíamos mencionado anteriormente, el número promedio de atractores crece con el tamaño de la red. Por lo tanto, es probable que la red mutada tenga más atractores que la red original. Así, además de los atractores originales, el paisaje epigenético de la red mutada puede contener nuevos atractores, que corresponderían a nuevas fenotipos en el organismo. En la Fig. 27 mostramos la probabilidad de que la red mutada, además de tener exactamente los mismos atractores que la red original, tenga más. En otras palabras, esta es la probabilidad de desarrollar nuevos fenotipos sin cambiar los ya existentes. La Fig. 27 se muestra dicha probabilidad para redes operando en las tres fases dinámicas, y como vemos, la probabilidad es máxima para redes críticas.

El resultado mostrado en la Fig. 27 es de la mayor importancia. No sólo estamos pidiendo que la red sea robusta al conservar todos sus atractores, sino que además estamos pidiendo que sea evolucionable ya que nuevos atractores emergen. Esta coexistencia entre robustez y evolucionabilidad hace posible que los organismos reutilicen las soluciones que se han desarrollado para adaptarse a diversos retos ambientales, y que al mismo tiempo sean capaces de generar nuevas soluciones para adaptarse a nuevos retos. Los organismos no se reinventan en cada paso evolutivo, sino que utilizan las soluciones ya existentes y generan nuevas. Y esta coexistencia entre robustez (mantener las soluciones ya existentes) y evolucionabilidad (generar nuevas soluciones) ocurre con la mayor probabilidad en redes con dinámicas críticas. Una representación visual de dicha coexistencia aparece en la Fig. 28.

12.7. Importancia de la criticalidad genética

Utilizando los datos experimentales más completos que se han producido hasta el momento en cuanto a la estructura y expresión genómicos, hemos demostrado que las redes genéticas de cinco organismos de reinos diferentes exhiben dinámicas críticas. El hecho de que estas redes muestren todas criticalidad dinámica sugiere fuertemente que la criticalidad es una característica genérica que ha surgido en la evolución porque de alguna manera confiere ventajas adaptativas a los organismos. En particular, estudiando el comportamiento del paisaje de atractores en poblaciones de redes Booleanas sujetas mutaciones, hemos encontrado que es justo en la fase crítica donde coexisten con la máxima probabilidad la robustez y la evolucionabilidad ob-

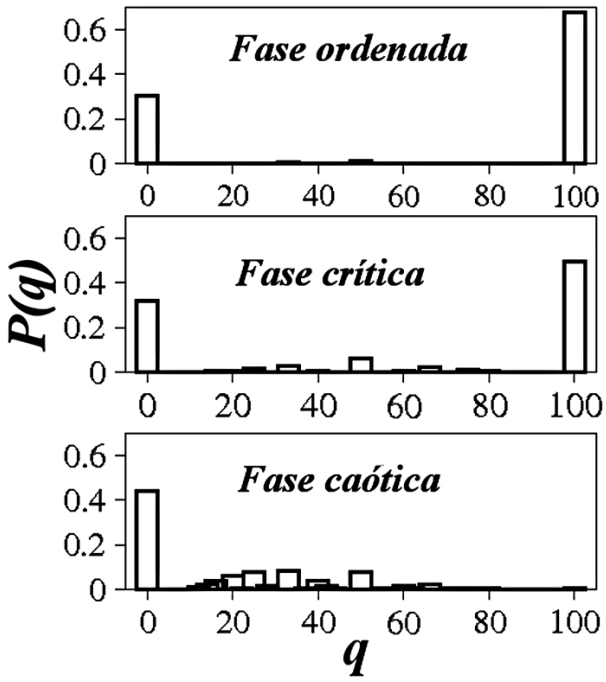


Figura 26: Probabilidad $P(q)$ de que se conserve un porcentaje q de los atractores originales después de la duplicación y divergencia de un solo gen en la red de Kauffman. En las fases ordenada y crítica ocurre con la máxima probabilidad que todos los atractores originales se conservan, mientras que en la fase caótica lo más probable es que ninguno de los atractores originales se conserva.

servadas en los organismos vivos. Esta coexistencia hace posible, por un lado, que los organismos conserven las características fenotípicas ya adquiridas para adaptarse a sus entornos, y por otro lado les da la plasticidad necesaria para desarrollar nuevas características. Las nuevas características adquiridas con la emergencia de nuevos atractores en el paisaje epigenético pueden ser favorables (o no) para que el organismo se enfrente a nuevos desafíos evolutivos. En cualquier caso, son precisamente las características del fenotipo que emergen con los nuevos atractores dinámicos de la red, las que fungen como el material sobre el cual actúa la selección natural para general organismos adaptados a sus entornos. Es importante notar la diferencia entre la nueva función que adquiere el gen duplicado después de la divergencia, y los nuevos fenotipos que adquiere toda la red. En la literatura científica generalmente

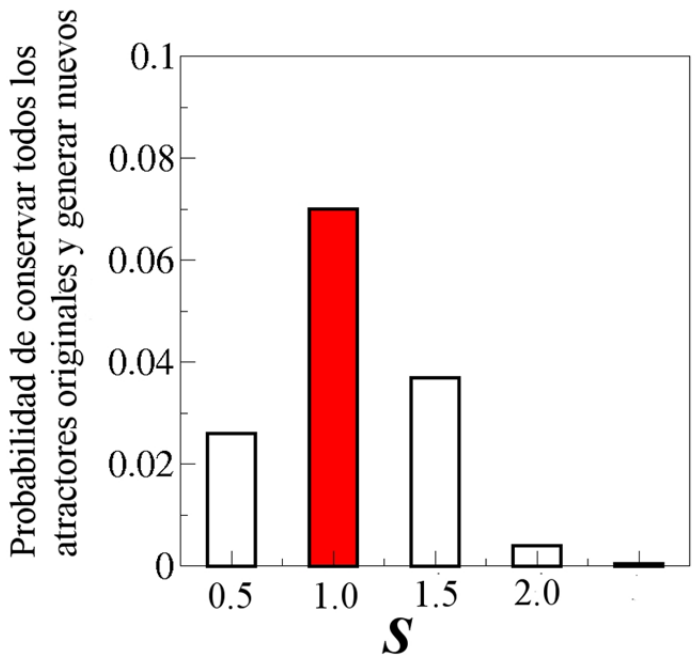


Figura 27: Probabilidad de que la red mutada conserve todos los atractores de la red original y que al mismo tiempo tenga nuevos atractores, como función de la susceptibilidad de la red. Nótese que es para la fase crítica ($S = 1$) para la que se tiene la máxima probabilidad de coexistencia entre robustez y evolucionabilidad.

se dice que la divergencia genética hace que el gen duplicado adquiera una nueva función (puede generar una nueva proteína, o regular a un conjunto diferente de genes, etc.). Sin embargo, hemos visto que este tipo de "selección relajada", en el cual el gen duplicado se muta y se retiene, puede cambiar la función de la red genética completa, es decir, puede generar atractores nuevos en los que se codifican nuevas características fenotípicas de todo el organismo. Esto puede ser el mecanismo por el cual ocurre el "equilibrio puntuado" en la evolución, proceso propuesto por N. Eldredge y S.J. Gould que establece que muy frecuentemente los cambios evolutivos no se dan de forma gradual, poco a poco, sino que ocurren en eventos rápidos que generan ramificación en las especies.

Así, la duplicación y divergencia genética es un mecanismo que constantemente está reestructurando el paisaje epigenético, expandiéndolo y explo-

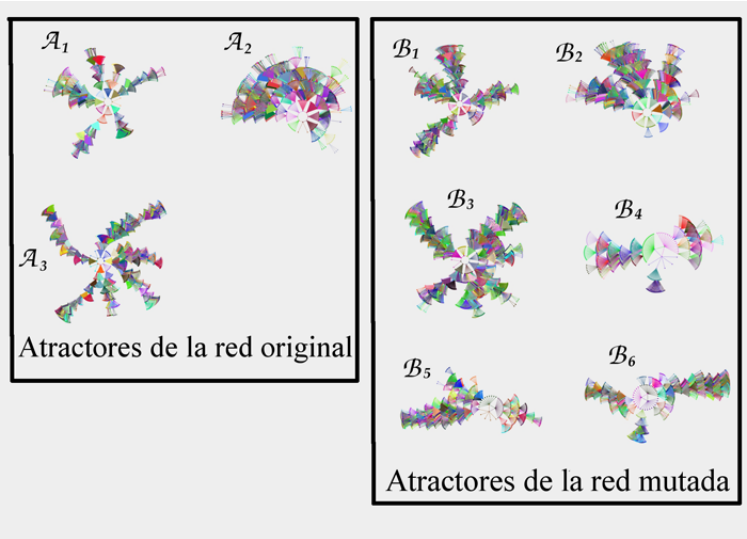


Figura 28: Coexistencia de robustez y evolucionabilidad. El cuadro de la izquierda muestra el paisaje de atractores de la red original, mientras que el cuadro de la derecha muestra el correspondiente paisaje de la red mutada. Los atractores $\mathcal{B}_1, \mathcal{B}_2$ y $\mathcal{B}_3$ son idénticos a los atractores $\mathcal{A}_1, \mathcal{A}_2$ y $\mathcal{A}_3$, por lo que el fenotipo codificado en dichos atractores se conserva después de la duplicación y divergencia. Sin embargo, nuevos atractores emergen ($\mathcal{B}_4, \mathcal{B}_5$ y $\mathcal{B}_6$) en el paisaje de la red mutada, lo cual le confiere al organismo la posibilidad de evolucionar.

rando nuevas soluciones ante retos ambientales cambiantes. La criticalidad dinámica hace posible que esta exploración y reestructuración del paisaje epigenético se lleve a cabo de forma robusta, sin que el organismo pierda las características ya adquiridas, generándose así la gran variedad de funcionamientos estables que observamos en la naturaleza que nos rodea.

13. Epílogo

La teoría de redes complejas es un campo de gran actividad científica en la actualidad. Basta con hacer una búsqueda en Google con la frase "Complex Networks" para darnos cuenta de la gran cantidad de investigadores que se encuentran trabajando en este campo. En estas notas introductorias hemos visto sólo *algunas* de las propiedades importantes de las redes complejas, pero la teoría no se termina aquí. Actualmente uno de los principales

desafios es caracterizar las propiedades dinámicas de las redes complejas dado que conocemos su topología. Por ejemplo, la estructura de la red de regulación genética de *Escherichia coli* es conocida. Sin embargo, aún no podemos predecir los diferentes fenotípos que resultan de la dinámica de dicha red. Probablemente la persona que resuelva este problema se ganará el premio Nobel. Pero lo que nos interesa ultimadamente es encontrar las leyes de organización de la materia, si es que tales leyes existen. El hecho de que redes tan diferentes presenten el mismo tipo de topología libre de escala sugiere que puede existir una ley fundamental que determine este tipo de estructuras. Igualmente, el hecho de que especies de organismos tan distintas como las bacterias, los insectos y las plantas hayan evolucionado hacia dinámicas críticas en la red genética sugiere que existen leyes dinámicas que determinan las características fenotípicas de los organismos. Dichas leyes no se han encontrado aún y tal vez nunca se encuentren, pero el viaje hacia su descubrimiento ha resultado más que divertido. Espero con estas notas haberlos motivado para que se unan al estudio del fascinante mundo de las redes complejas y a la búsqueda de las leyes de organización de la materia. ¡En hora buena!

Apéndices

A. Distribución de Poisson como límite de una distribución binomial

En este apéndice mostramos cómo obtener la Eq. (6) a partir de la Eq. (5), tomando el límite $N \to \infty$ y $M \to \infty$. Para esto utilizaremos dos resultados bien conocidos: la fórmula de Stirling y la definición de e^a:

$$n! \approx e^{-n} n^n \text{ para } n \text{ grande,}$$

$$e^a = \lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n$$

Utilizando la fórmula de Stirling, el coeficiente binomial $\binom{N-1}{k}$ que aparece en la ecuación (5) puede escribirse, para N grande, como

$$\begin{aligned}
\binom{N-1}{k} &= \frac{(N-1)!}{k!(N-1-k)!} \\
&\approx \frac{e^{-(N-1)}(N-1)^{N-1}}{k!\; e^{-(N-1-k)}(N-1-k)^{N-1-k}} \\
&= \frac{e^{-k}(N-1-k)^k}{k!} \left(\frac{N-1}{N-1-k}\right)^{N-1}
\end{aligned}$$

Por lo tanto, la distribución binomial (5) queda, para N grande, como

$$P(k) = \frac{e^{-k}(N-1-k)^k}{k!} \left(\frac{N-1}{N-1-k}\right)^{N-1} (p_e)^k \, (1-p_e)^{N-1} \tag{29}$$

Para continuar, notemos que la cantidad $z = 2M/N$ es la conectividad promedio de cada nodo en la red. En términos de esta cantidad, la probabilidad p_e puede escribirse como $p_e = z/(N-1)$, con lo cual la ecuación (29) queda como

$$P(k) = \frac{e^{-k}(N-1-k)^k}{k!} \left(\frac{N-1}{N-1-k}\right)^{N-1} \left(\frac{z}{N-1}\right)^k \left(1 - \frac{z}{N-1}\right)^{N-1}$$

lo cal, después de arreglar algunos términos, puede escribirse como

$$P(k) = \frac{e^{-k}}{k!} z^k \left(1 - \frac{k}{N-1}\right)^k \left(\frac{1 - z/(N-1)}{1 - k/(N-1)}\right)^{N-1}$$

tomando el límite $N \to \infty$ y $M \to \infty$ de tal forma que la conectividad promedio $z = 2M/N$ permanezca constante, y utilizando el hecho de que $e^a = \lim_{n\to\infty}(1 + a/n)^n$, obtenemos finalmente

$$P(k) = e^{-z}\ \frac{z^k}{k!}$$

Por lo tanto, para una red muy grande, el proceso de "hilvanar" parejas de botones escogidas al azar genera una distribución de conexiones de Poisson.

B. Ecuación Maestra

Supongamos que tenemos un sistema que puede estar en cualquiera de los estados E_1, E_2, ..., E_N. Sea $W_{m\to n}(t)$ la probabilidad condicional de que el sistema, dado que estaba en el estado E_m al tiempo t, "brinque" al estado E_n. Las probabilidades $W_{m\to n}(t)$ se denominan *probabilidades de transición.*

Sea $P(n,t)$ la probabilidad de que el sistema se encuentre en el estado E_n al tiempo t. Esta probabilidad puede cambiar en el tiempo por dos factores:

1. El sistema estaba en el estado E_n al tiempo t y brincó a otro estado E_m, lo cual claramente disminuye la probabilidad $P(n, t+1)$ de encontrar al sistema en el estado E_n al tiempo $t+1$.

2. El sistema estaba en algún estado E_m al tiempo t y brincó al estado E_n, lo cual aumenta la probabilidad $P(n, t+1)$ de encontrar al sistema en el estado E_n al siguiente instante de tiempo.

Tomando en cuenta estos dos factores, la probabilidad de que el sistema se encuentre en el estado E_n al tiempo $t+1$ está dada por

$$P(n,t+1) = \sum_{m=1}^{N} P(m,t)W_{m\to n}(t) - \sum_{m=1}^{N} P(n,t)W_{n\to m}(t) \qquad (30)$$

El primer término del lado derecho nos dice que al tiempo t el sistema pudo estar en el estado E_m y brincar al estado E_n, aumentando así la probabilidad de estar en E_n al tiempo $t+1$. El segundo término toma en cuenta el hecho de que al tiempo t el sistema pudo estar en el estado E_n y salirse de allí, brincando al estado E_m y disminuyendo la probabilidad de estar en E_n al tiempo $t+1$. Las sumas sobre m toman en cuenta todos los posibles estados hacia o desde los que puede brincar el sistema.

La ecuación (30) es la ecuación maestra del sistema. El objetivo es resolverla para encontrar $P(n,t)$. Si las probabilidades de transición $W_{m\to n}$ son independientes del tiempo, la ecuación se puede resolver más o menos fácilmente. Pero la cosa se complica enormemente si las probabilidades de transición dependen del tiempo. Afortunadamente, las ecuaciones maestras que aparecen en la teoría de crecimiento de redes, al menos en los casos sencillos presentados en estas notas, son relativamente fáciles de resolver.

Agradecimientos

Este trabajo se realizó con el apoyo del proyecto PAPIIT-UNAM IN112407.

Referencias

[1] En la red del universo de Malvel, dos super héroes están conectados si han aparecido por lo menos una vez en el mismo comic. De este estudio, que puede encontrarse en http://dmi.uib.es/∼joe/marvel.html, resulta que es el hombre ara na el super héroe con el máximo número de conexiones.

[2] Statistical Mechanics of Complex Networks. Réka Albert y Albert-László Barabási. *Reviews of Modern Physics* **74**(1):47-97 (2002).

[3] The Structure and Function of Complex Networks. Mark E.J. Newman. *SIAM Review* **45**(2):167-256 (2003).

[4] Evolution of Networks. S.N. Dorogovtsev and J.F.F. Mendes. *Advances in Physics* **51**(4):1079-1187 (2002).

[5] Evolution of Networks: From Biological Nets to the Internet and WWW. S.N. Dorogovtsev and J.F.F. Mendes. Oxford University Press, Oxford. ISBN: 0198515901. (2003). Este libro puede encontrarse **GRATIS** en http://www.fyslab.hut.fi/∼sdo/

[6] Connectivity of Growing Random Networks. P. L. Krapivsky, S. Redner, y F. Leyvraz. *Phys. Rev. Lett.* **85**:4629-4632 (2000).

[7] Power-Law Distribution of the Word Wide Web. L. A. Adamic y B. Hubeerman. *Science* **287**(24):2115a (2000).

[8] Bose-Einstein condensation in complex networks. G. Bianconi y A.-L. Barabsi. *Phys. Rev. Lett.* **86**:5632-5635 (2001).

[9] Epidemic Spreading in Scale-Free Networks. Romualdo Pastor-Satorras y Alessandro Vespignani. *Phys. Rev. Lett.* **86**:3201-3203 (2001).

[10] Spread of Epidemic Disease on Networks. Mark E. J. Newman. *Phys. Rev. E.* **66**:16128 (2002).

[11] Dynamical Phase Transition in a Neural Network Model with Noise: an Exact Solution. Cristián Huepe-Minoletti and Maximino Aldana-Gonzalez. *Journal of Statistical Physics* **108**(3/4):527-540 (2002).

[12] Phase Transitions in Scale-Free Neural Networks: Departure from the Standard Mean-Field Universality Class. Maximino Aldana and Hernán Larralde. *Physical Review E* **70**:066130 (2004).

[13] Boolean dynamics of networks with scale-free topology. Aldana M. *Physica D* **185** (2003) 4566

[14] Robustness and evolvability in genetic regulatory networks. Aldana, M, Balleza E, Kauffman S y Resendiz O. *Journal of Theoretical Biology* **245**: 433-448 (2007).

[15] Critical dynamics in genetic regulatory networks: Examples from four kingdoms. Balleza E, Alvarez-Buylla ER, Chaos A, Kauffman S, Shmulevich I y Aldana M. *PLoS ONE* **3**(6): e2456 (2998).

Introducción a las resonancias de Feshbach

R. Cabrera-Trujillo
Instituto de Ciencias Físicas, UNAM

1 Introducción

El impacto que han tenido los gases atómicos y moleculares ultra-frios en la física atómica, molecular y óptica (FAMO) está ligado al extraordinario control que de estos sistemas se tiene para investigar el comportamiento fundamental de la materia cuántica bajo condiciones extremas. El interés va más allá de FAMO, alcanzando campos tales como materia condensada y física de pocos y muchos cuerpos. En todas estas aplicaciones, las resonancias de Feshbach representan la herramienta escencial para controlar las interacciones entre átomos, lo cual ha sido la clave para llevar a cabo muchos avances.

En estas notas, intentare dar una revisión sobre que son las resonancias de Feshbach. Para una revisión extensiva sobre el tema recomiendo el trabajo de Chin *et al.* [1] y las referencias ahí incluidas.

Los orígenes físicos y las propiedades elementales de una resonancia de Feshbach se pueden entender de una manera muy simple. Consideremos dos curvas de potencial moleculares $V_{bg}(R)$ y $V_c(R)$, como se muestran en la Fig. 1. Para grandes distancias de interacción R, el potencial de fondo $V_{bg}(R)$ conecta asimptoticamente a dos átomos libres en el gas ultra-frio. Para un proceso de colisión a bajas energías E, el potencial representa un canal abierto energéticamente, el cual llamaremos el canal de entrada. El potencial $V_c(R)$ representa un canal cerrado, el cual es importante puesto que puede tener un estado molecular ligado cerca del umbral del canal abierto.

Una resoncia de Feshbach ocurre cuando el estado molecular ligado en el canal cerrado energéticamente se acerca al estado dispersivo en el canal cerrado. Entonces, acoplamiento, aunque sea debil, conduce a un mezclado fuerte entre los dos canales. La diferencia en energía se puede controlar mediante un campo megnético, cuando los momentos magnéticos correspondientes son diferentes. Esto se llama una resonancia de Feshbach magnéticamente sintonizable. Alternativamente, acoplamiento resonante se puede lograr mediante métodos ópticos, lo cual se llama una resoncia de Feshbach óptica o de Fano.

Una resoncia de Feshbach magnéticamente sintonizable se expresa como

$$a(B) = a_{bg}\left(1 - \frac{\Delta}{B - B_0}\right) \tag{1}$$

que fue primero derivada por Moerdijk [8] para una longitud de dispersión de onda-s, a, como función del campo magnético B. La longitud de dispersión de

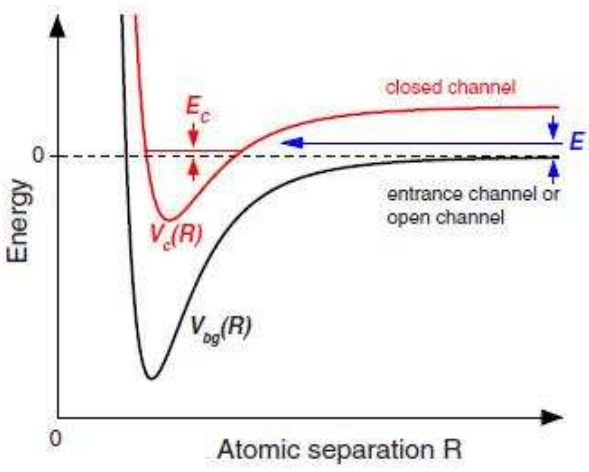


Figure 1: Modelo básico de dos canales para un resonancia de Feshbach. El fenómeno ocurre cuando dos átomos colisionando con energía E en el canal de entrada se acoplan resonantemente a un estado ligado molecular E_c que está en el canal cerrado.

fondo a_{bg} asociada al potencial $V_{bg}(R)$ representa el valor fuera de resonancia. El parámetro B_0 denota la posición de la resonancia que es donde la longitud de dispersión diverge ($a \to \infty$) y el parámetro Δ es el ancho de la resonancia. Notemos que a_{bg} y Δ pueden ser negativos o positivos.

La Fig. 2 muestra una observación experimental de una resonancia de Feshbach [7] en la que se muestran dos características importantes de una resonancia de Feshbach.

1. La longitud de dispersión es sintonizable.

2. La pérdida de átomos en el gas ultra-frio en la región resonante.

Esto último se atribuye al fuerte incremento de la recombinación de tres cuerpos y a la formación de moléculas como resultado de la resonancia de Feshbach.

Las primeras investigaciones de fenómenos que surgieron del acoplamiento entre un estado ligado y el continuo se llevaron a cabo en la década de los 30's del siglo pasado. Rice [9] estudio como un estado ligado se pre-disocia en el continuo. Fano [3] estudio perfiles asimétricos de lineas como resulado de interferencia cuántica. Fano [3] y Feshbach [6] desarrollaron de forma independiente sus tratamientos al fenómeno resonante. El trabajo de Feshbach se desarrolló en el contexto de la física nuclear mientras que el trabajo de Fano tuvo como fondo la física atómica. De ahí que algunas veces se use el término resonancias de Fano-Feshbach.

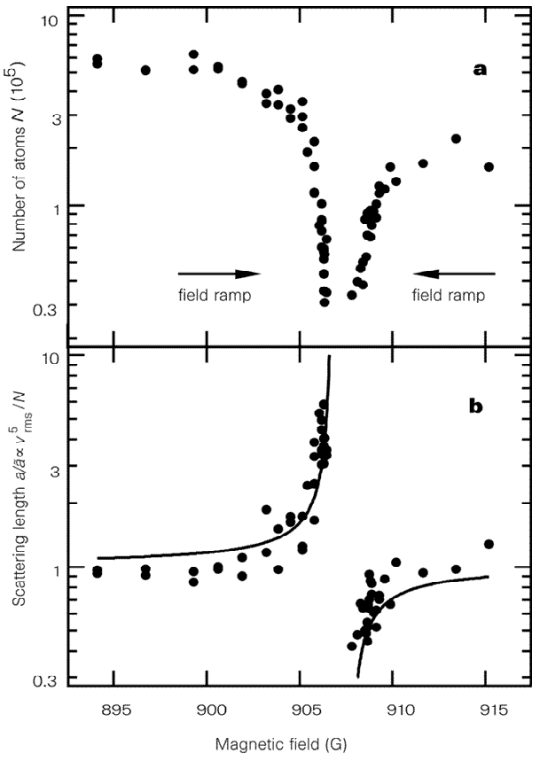


Figure 2: Observación de una resonancia de Feshbach magnéticamente sintonizable en un condensado de Bose-Einstein ópticamente atrapado de átomos de Sodio. El panel superior muestra una fuerte pérdida de átomos cerca de la resonancia, la cual es debida al incremento de la recombinación de tres cuerpos. El panel inferior muestra la forma dispersiva de la longitud de dispersión a cerca de la resonancia. Figura tomada de Inouye *et el.* [7].

2 Revisión de teoría de dispersión

En esta sección daremos un breve resumen de los elementos básicos de teoría de dispersión para colisiones en un solo canal e introduciremos la longitud de dispersión, la cual caracteriza interacciones a bajas energías en un par de parículas.

Consideremos la dispersión de dos partículas de masa m_1 y m_2, las cuales por el momento supondremos que no tienen grados de libertad internos y asumiremos que son indistinguibles. Como es usual, transformaremos al centro de masa del sistema y a coordenadas relativas. La función de onda para el movimiento del centro de masa es una onda plana, mientras que para el movimiento relativo se satisface la ecuación de Schrödinger para una partícula de masa reducida $\mu = m_1 m_2/(m_1 + m_2)$ entre dos partículas. Para describir la dispersión, escribimos la función de onda para el movimiento relativo como la suma de una onda

plana incidente y una dispersada

$$\psi = e^{ikz} + \psi_{sc}(\mathbf{r}) \tag{2}$$

donde hemos escogido la velocidad relativa de la onda incidente en la dirección z. Para separaciones interatómicas grandes, la onda dispersada es una onda esférica saliente $f(\mathbf{k})\exp(ikr)/r$, donde $f(\mathbf{k})$ es la amplitud de dispersión y $\mathbf{k}$ es el vector de onda de la onda dispersada. Asumiendo que la dispersión entre átomos es simétricamente esférica, entonces $f(\theta)$ que dependen solo de la dirección a travez del ángulo de dispersión θ, el cual es el ángulo entre la dirección del momento relativo de los átomos antes y después de la colisión. Entonces

$$\psi = e^{ikz} + f(\theta)\frac{e^{ikr}}{r} \tag{3}$$

La energía del sistema es $E = \hbar^2 k^2/2\mu$. Para bajas energías, es suficiente considerar dispersión de ondas-s, como veremos mas adelante. En este límite ($k \to 0$) la amplitud de dispersión $f(\theta)$ se reduce a una constante, la cual denotaremos como $-a$. Así

$$\psi = 1 - \frac{a}{r} \tag{4}$$

La constante a se llama longitud de dispersión. Ahora veremos como podemos relacionar a a con el potencial de la interacción.

La sección diferencial de dispersión, *i.e.* la sección transversal de colisón por unidad de ángulo sólido está dada por

$$\frac{d\sigma}{d\Omega} = |f(\theta)|^2 \tag{5}$$

Para dispersión en un ángulo entre θ y $\theta + d\theta$, el elemento de ángulo solido es $d\Omega = 2\pi \sin\theta d\theta$. Como el potencial es simétricamente esférico, la solución tiene simetría axial respecto a la dirección de incidencia. La funcion de onda se puede expander en términos de polinomios de Legendre

$$\psi = \sum_{l=0}^{\infty} A_l P_l(\cos(\theta)) R_{kl}(r) \tag{6}$$

donde $R_{kl}(r)$ satisface

$$\frac{d^2 R_{kl}(r)}{dr^2} + \frac{2}{r}\frac{dR_{kl}(r)}{dr} + \left[k^2 - \frac{l(l+1)}{r^2} - \frac{2\mu}{\hbar^2}V(r)\right] R_{kl}(r) = 0 \tag{7}$$

donde $V(r)$ es el potencial de interacción. Para $r \to \infty$ la funcion de onda radial es

$$R_{kl} \sim \frac{1}{kr}\sin(kr - \frac{\pi l}{2} + \delta_l)\,. \tag{8}$$

Aquí δ_l es el corrimiento de fase debido al potencial. Comparando estas ecuaciones y expandiendo la onda plana en polinomios de Legendre uno encuentra

que $A_l = i^l(2l+1)\exp(i\delta_l)$ y

$$f(\theta) = \frac{1}{2ik}\sum_{l=0}^{\infty}(2l+1)(e^{2\delta_l}-1)P_l(\cos(\theta)) \tag{9}$$

tal que la sección transversal es

$$\sigma = \frac{4\pi}{k^2}\sum_{l=0}^{\infty}(2l+1)\sin^2(\delta_l)\,. \tag{10}$$

La seccion transversal está dominada por el término $l=0$, el cual se denomina dispersión de onda s correspondiente a una amplitud de dispersión $f = \delta_0/k$ con función de onda radial

$$R_{k0} \sim c_1\frac{\sin kr}{kr} + c_2\frac{\cos kr}{r} \tag{11}$$

tal que

$$\delta_0 = k\frac{c_2}{c_1} \tag{12}$$

De la definición de longitud de dispersión en términos de la función de onda para $k \to 0$, uno encuentra que

$$\delta_0 = -ka \tag{13}$$

lo cual muestra que a está determinada por los coeficientes en la solución asintótica

$$a = -\left.\frac{c_2}{c_1}\right|_{k\to 0} \tag{14}$$

En el límite $k \to 0$ entonces

$$\sigma = \frac{4\pi}{k^2}\delta_0^2 = 4\pi a^2 \tag{15}$$

Así, a bajas energías, la longitud de dispersión determina la dispersión del sistema.

En la aproximación de Born, la longitud de dispersión está dada por

$$a_{Born} = \frac{\mu}{4\pi\hbar^2}\int V(\mathbf{r})d^3\mathbf{r} \tag{16}$$

correspondiente a $|\mathbf{k}-\mathbf{k}'| = 0$. Dependiendo del potencial, la longitud de dispersión puede tomar cualquier valor entre $-\infty < a < \infty$.

3 Resonancias de Feshbach

Ahora, la energía del canal $E_\alpha = E(q_1) + E(q_2)$ es la suma de las energías de los átomos separados. Asumiendo que los átomos están preparados en un

canal α con energía cinética relativa E, tal que la energía total del sistema es $E_T = E_\alpha + E$. Cualquier canal β con $E_\beta < E_T$ se llama un canal abierto y cualquier canal con $E_\beta > E_T$ será un canal cerrado. Una colisión puede producir átomos en un canal abierto después de la colisión, pero no en un canal cerrado, puesto que los átomos no tienen suficiente energía para separarse a los átomos producto.

Una resonancia "convencional" ocurre cuando el corrimiento de fase δ_l cambia rápidamente en π sobre un rango pequeño de energías debido a la presencia de un estado cuasi-ligado del sistema que se acopla a un estado de dispersión de los átomos que colisionan. Este estado cuasi-ligado puede estar atras de una barrera repulsiva de un potencial, la cual define una resonancia de forma, o puede deberse a estados con diferente simetría o potencial del de los átomos que colisionan, entonces se llama resonancia de Feshbach.

Consideremos un hamiltoniano H de dos canales que describe a nuestro sistema como una buena aproximación a un problema de dos canales desacoplados. Uno es el canal de fondo dispersivo $|bg\rangle$ con estado dispersivo $|E\rangle = \phi_{bg}(R, E)|bg\rangle$ etiquetado por su energía de colisión E. El otro estado es el canal cerrado $|c\rangle$ que tiene un estado ligado $|C\rangle = \phi_c(R)|c\rangle$ con energí E_c. Las funciones $\phi_c(R)$ y $\phi_{bg}(R, E)$ son soluciones a la ecuación de Schrödinger radial (ec. 7). Aquí $\phi_c(R)$ está normalizada a la unidad. La dispersión en el canal abierto está caracterizada por el corrimiento de fase de fondo $\delta_{bg}(E)$. Cuando el acoplamiento $W(R)$ entre los dos canales se toma en cuenta, entonces los dos canales se mezclan por la interacción y la dispersión toma un término resonante extra debido al estado ligado embebido en el continuo de la dispersión

$$\delta(E) = \delta_{bg}(E) + \delta_{res}(E) \tag{17}$$

donde $\delta_{res}(E)$ está dada por

$$\tan \delta_{res}(E) = -\frac{\frac{1}{2}\Gamma(E_c)}{E - E_c - \delta E(E_c)} \tag{18}$$

La interacción $W(R)$, la cual se anula para R grandes, determina dos características de la resonancia: su ancho

$$\Gamma(E) = 2\pi \left|\langle C|W(R)|E\rangle\right|^2 \tag{19}$$

y el corrimiento a una nueva posición

$$\delta E(E) = \mathcal{P} \int_{-\infty}^{\infty} \frac{\left|\langle C|W(R)|E'\rangle\right|^2}{E - E'} dE' \tag{20}$$

donde $\mathcal{P}$ es la parte principal de la integral. Tomemos como aproximación que el ancho y el corrimiento son constantes independientes de la energía, *i.e.* $\Gamma(E_c)$ y $\delta E(E_c)$ evaluadas en al energía resonante E_c. La fase resonante cambia en π cuando E cambia en un rango del ancho de Γ. Cuando $k \to 0$ se tiene que

$$\frac{1}{2}\Gamma(E) \to (k a_{bg})\Gamma_0 \tag{21}$$

y

$$E_c + \delta E(E) \to E_0 \tag{22}$$

donde Γ_0 y E_0 son constantes independientes de E. Así

$$-ka = -ka_{bg} - \frac{ka_{bg}\Gamma_0}{-E_0 + i\gamma/2} \tag{23}$$

donde hemos agregado, por generalidad, una razón de decaimiento $\gamma/\hbar$ para el decaimiento del estado ligado. Así, la habilidad para sintonizar el umbral de la posición de la resonancia E_0 que pasa por cero es lo que determina la posición resonante.

En el caso de resonancias sintonizables magnéticamente existe una diferencia en el momento magnético $\delta\mu_{mag} = \mu_{atomo} - \mu_0$ entre en momento magnético de los átomos μ_{atomos} y el momento magnético μ_0 de los estados ligados para $|C\rangle$. Así la energí E_0 del estado $|C\rangle$ relativa al canal de energí de los átomos separados es

$$E_c = \delta\mu_{mag}(B - B_0) \tag{24}$$

y dado que $\gamma = 0$, entonces

$$a = a_{bg}\left(1 - \frac{\Delta}{B - B_0}\right) \tag{25}$$

donde $\Delta = \Gamma_0/\delta\mu_{mag}$ y $B_0 = B_c + \delta B$ son el ancho y posición de la singularidad en la longitud de dispersión, recorrida debido a la interacción entre los estado cerrados y abiertos.

Así, la determinación de una resonancia de Feshbach se reduce a conocer de manera precisa las curvas de potencial para un sistema multi-electrónico.

En mi grupo de investigación estamos interesados en el estudio de ondas de materia y su propagación en chips atómicos [4, 5] lo cual requiere el conocimiento de la longitud de dispersión del sistema la cual aparece en la ecuación de Gross-Pitaevskii [2] y que determina la dinámica de la onda de materia.

4 Agradecimientos

Este trabajo ha sido apoyado con el proyecto de investigación PAPIIT IN-101-611 de la UNAM.

References

[1] Cheng Chin, Rudolf Grimm, Paul Julienne, and Eite Tiesinga. Feshbach resonances in ultracold gases. *Rev. Mod. Phys.*, 82:1225–1286, Apr 2010.

[2] Franco Dalfovo, Stefano Giorgini, Lev P. Pitaevskii, and Sandro Stringari. Theory of bose-einstein condensation in trapped gases. *Rev. Mod. Phys.*, 71:463–512, Apr 1999.

[3] U. Fano. Sullo spettro di assorbimento dei gas nobili presso il limite dello spettro d'arco. *Nuovo Cimento*, 12:154, 1935.

[4] József Fortágh and Claus Zimmermann. Toward atom chips. *Science*, 307:860, 2005.

[5] József Fortágh and Claus Zimmermann. Magnetic microtraps for ultracold atoms. *Rev. Mod. Phys.*, 79(1):235–289, Feb 2007.

[6] Feshbach H. Unified theory of nuclear reactions. *Ann. Phys. (N. Y.)*, 5:357, 1958.

[7] S. Inouye, M. R. Andrews, J. Stenger, H.-J. Miesner, D. M. Stamper-Kurn, and W. Ketterle. Observation of feshbach resonances in a bose-einstein condensate. *Nature*, 392:151 – 154, 1998.

[8] A. J. Moerdijk, B. J. Verhaar, and A. Axelsson. Resonances in ultracold collisions of ^{6}Li, ^{7}Li, and ^{23}Na. *Phys. Rev. A*, 51:4852–4861, Jun 1995.

[9] O. K. Rice. Predissociation and the crossing of molecular potential energy curves. *J. Chem. Phys.*, 1:375, 1933.

PLASMAS

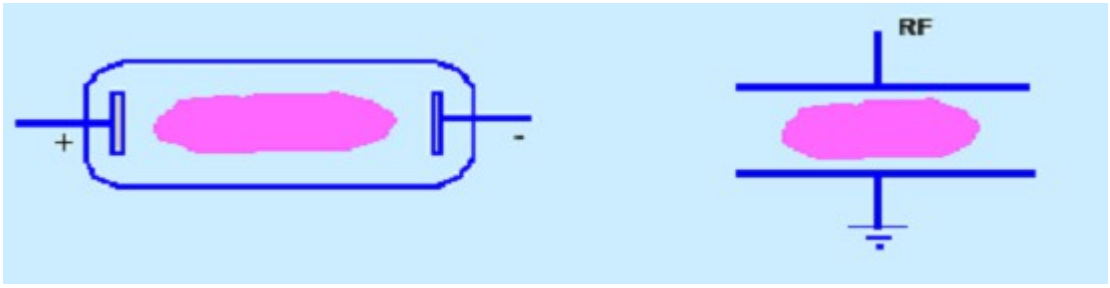
B. Campillo, O. Flores y H. Martínez.
Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Apartado Postal 48-3, 62191 Cuernavaca, Morelos, México. hm@fis.unam.mx

1.- Introducción.

Una de las técnicas más comunes para ionizar un gas es la descarga eléctrica entre dos electrodos. Las descargas eléctricas en gases ocurren como consecuencia de la ionización del medio y aparecen muy a menudo de forma natural (rayos, chispas, arcos eléctricos, etc.) y también en muchas aplicaciones tecnológicas (lámparas de Ne, equipos de soldadura eléctrica, etc.).

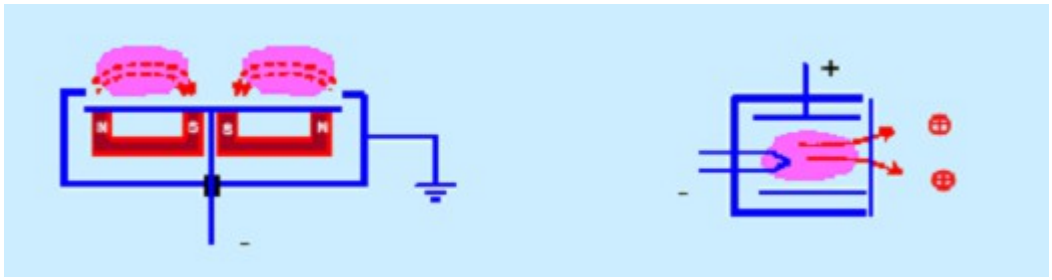
En todas estas aplicaciones se trata de obtener ventajas de los fenómenos asociados a la ionización de gases, y en particular del alto grado de excitación que alcanzan las moléculas ionizadas. Así, las moléculas, cuando se des-excitan, pueden desprender energía en forma de luz o de calor que es aprovechable en otros procesos.
Actualmente, las descargas eléctricas se utilizan mucho en el procesado de materiales. En este capítulo se verán aquellos aspectos de las descargas eléctricas directamente relacionados con la preparación de capas delgadas.

•Ejemplos de aplicación de las descargas eléctricas:



Descarga luminosa en un tubo a través de un gas enrarecido

Reactor de plasma-CVD para depósito (o ataque) de capas delgadas



Sputtering magnetrón (tipo planar)

Fuente de iones RF

Figura 1. Tipos de sistemas para generación de plasmas.

2. Plasmas

¿Qué es un plasma? Es un gas de partículas formado por:
–especies cargadas (electrones y iones).
–especies neutras (átomos o moléculas)
–radicales atómicos o moleculares y otras especies excitadas.

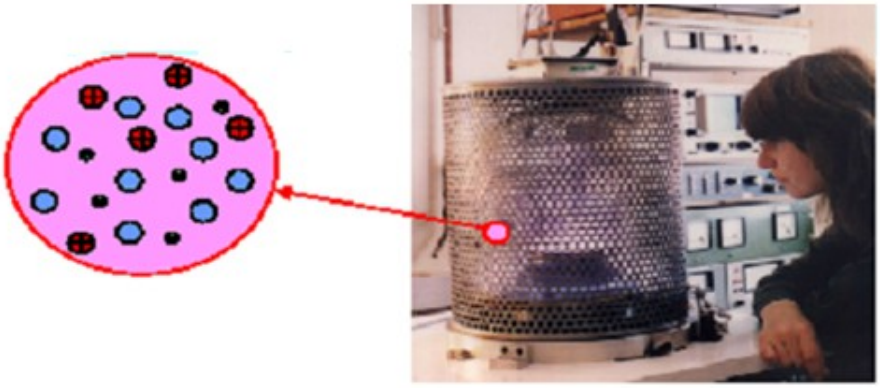


Figura 2. Cámara de descarga.

2.1.- Características:

• Conjunto de cargas neutro, con una distribución homogénea. Cualquier descompensación de carga en un punto produciría campos eléctricos tendentes a eliminar la descompensación.

• Al mismo tiempo, todo el conjunto de especies cargadas en el plasma tiene un comportamiento colectivo, al cual es posible asignar un potencial, frecuencia de vibración, energía media, etc.

•La existencia de especies excitadas da lugar a fenómenos de emisión luminosa originados por procesos continuos de excitación y des-excitación.

2.2- Grado de ionización (i) (Con $n_i = n_e$).

Dos casos extremos:

$$i = \frac{n_i}{n_i + n_0}$$

- i <<1 → plasma 'débilmente ionizado', (propiedades determinadas fundamentalmente por la presencia de las especies neutras).
- i ≈1 → plasma 'altamente ionizados' (las partículas neutras tienen un papel menor)

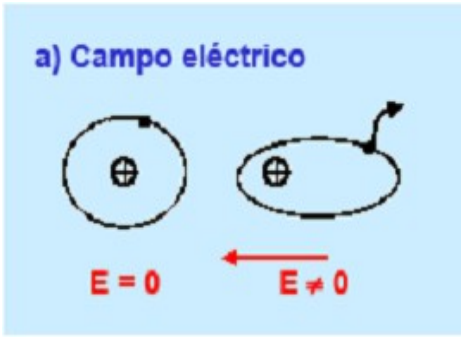
Notar: La presencia de cargas móviles en el plasma hace que el medio se comporte como un conductor, es decir con resistencia eléctrica prácticamente nula.

2.3.- Mecanismos de ionización:

a) Campo eléctrico.
b) Temperatura, colisiones múltiples entre partículas (átomos neutros, electrones, fotones, etc).
c) Impacto directo: átomo, fotón, etc., con alta energía.

2.4.- Densidad del plasma:

– Concentración de partículas cargadas/volumen, ρ
– Rango típico: $\rho = 10^{11}$-10^{19} cm^{-3}.
Descargas luminosas: $\rho \approx 10^{12}$ cm^{3}
Descargas en arco: $\rho \approx 10^{16}$ cm^{-3}
Recordar:
Para un gas a 1 atmósfera, $n = 2{,}46 \times 10^{19}$ cm^{-3}.
En situación de equilibrio: ρ constante.
Se trata de un equilibrio dinámico, en un proceso continuo de generación y pérdida de carga a través de colisiones.



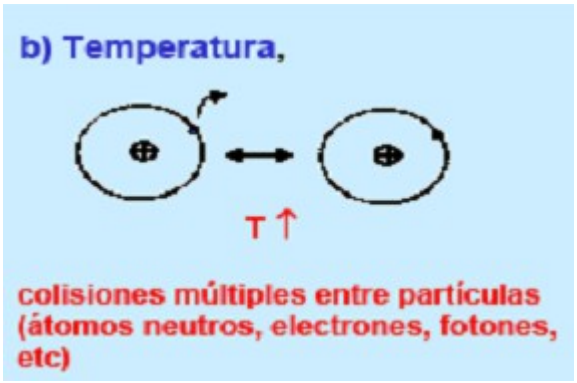


Figura 3. Mecanismos de ionización.

2.5.-Tipos de colisión:
– Elásticas: Se redistribuye el momento entre las partículas y la energía permanece constante. Ejemplo:

$$e^-_{\text{rápido}} + A_{\text{lento}} \rightarrow e^-_{\text{menos rápido}} + A_{\text{menos lento}}$$

–Inelásticas: igual, pero una fracción de la energía cinética se transfiere como energía interna a la otra partícula:

$$e^-_{\text{rápido}} + A \rightarrow e^-_{\text{lento}} + A^*$$
$$\rightarrow e^-_{\text{lento}} + A^+ + e^-$$

–Parámetros característicos de las colisiones:
- Energía cinética de las partículas, E_c.
- Energía de ionización, E_i.
- Recorrido libre medio, $\lambda = kT/(P\pi\sigma^2 2^{1/2})$
- Sección eficaz de colisión, σ: depende de la energía, E_c.

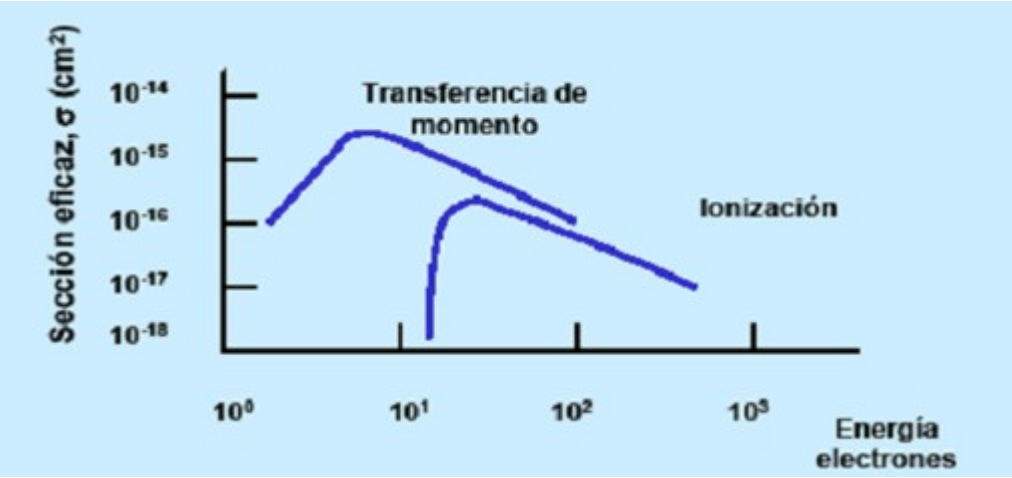


Figura 4. Sección eficaz de dos procesos de colisión.

2.6.- Energía de las partículas cargadas:
– Tanto las partículas excitadas del plasma (átomos, moléculas neutras, iones, etc.) así como los electrones están en movimiento aleatorio, con una cierta energía cinética media, etc.
- Dependiendo del grado de interacción y del proceso de aceleración, los electrones del plasma pueden tener una energía muy diferente a la de los iones y partículas neutras.
–La función de distribución de energía de los electrones, F(E) es compleja. En casos simples se asimila a veces a la del gas neutro (Maxwelliana) y por ello es posible asignar una temperatura equivalente de los electrones, T_e.
- En los plasmas utilizados en el procesado de materiales, la energía media de los electrones puede variar entre 1 y 20 eV, lo que implica temperaturas equivalentes de 10^4-10^5 K.

Notar: 1eV = 11600 K (Además, T_e puede ser muy diferente de la temperatura del gas).

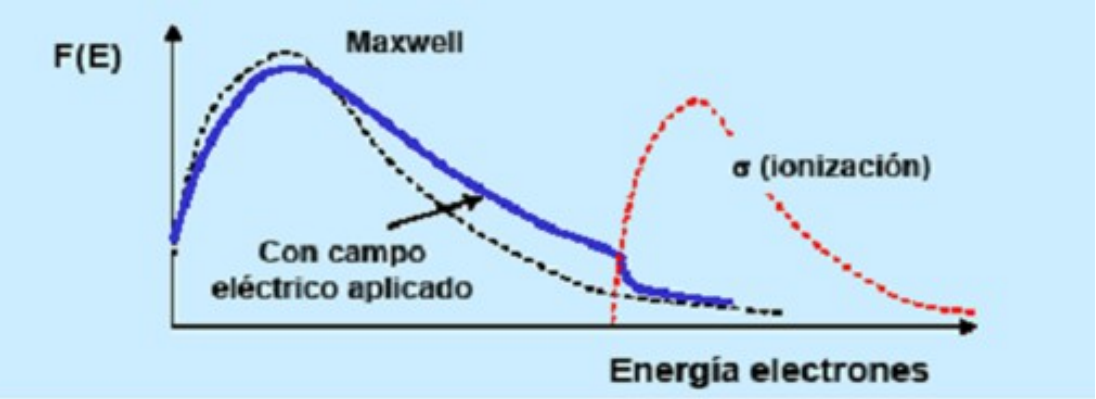


Figura 5. Distribución de energía.

2.7.- Neutralidad y límites del plasma:
– A nivel macroscópico el plasma es neutro, ya que $n_e = n_i$.
– Si por alguna razón se produce un exceso de carga (p. e. negativa) en un punto, ésta se rodea de carga positiva, hasta una cierta distancia, "longitud de Debye", a partir de la cual el campo eléctrico es cero:
λ_D= longitud de apantallamiento = $69 \cdot (T_e/n_e)$.
- Del mismo modo, un material en contacto o en el interior del plasma queda siempre cargado negativamente, ya que el flujo de electrones hacia el material se mayor que el de iones. Se forma así una doble capa, o funda del plasma, con el plasma cargado positivamente respecto al material:

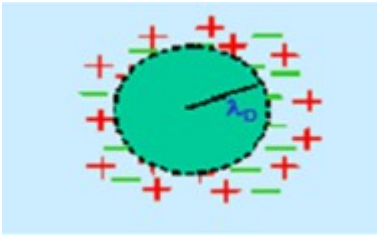


Figura 6. Longitud de Debye.

Notar: Cuando se aplica un campo eléctrico, las cargas del plasma se distribuyen hasta anular el campo en el interior (similar al caso de un material conductor).

2.8.- Descargas eléctricas entre dos electrodos.
• La mayoría de los plasmas utilizados en la preparación de capas delgadas mediante técnicas de PVD o CVD se generan por descargas eléctricas a presiones reducidas (plasmas fríos) o a la presión atmosférica (plasmas térmicos) utilizando dos electrodos polarizados en corriente continua (CC) o en corriente alterna (CA). En otros casos se utilizan también llamas procedentes de la combustión de gases (antorchas de proyección) o mediante radiación láser.

•En el caso de las descargas eléctricas las moléculas del gas se someten a un campo eléctrico suficientemente intenso para llevarlas a un estado de ionización con pérdida de uno o más electrones, quedando cargadas positivamente.
• Cuando se alcanza el estado estacionario existe un equilibrio dinámico entre las cargas que se pierden en las paredes y en los electrodos y las que se generan mediante fenómenos de ionización (descargas 'auto-mantenidas).
•Existe también un balance entre la energía ganada por los electrones y por los iones a partir del campo eléctrico y las pérdidas producidas en colisiones, sobre todo de tipo inelástico (de los electrones con especies neutras que son las más abundantes). Estas colisiones inelásticas son las que llevan a las especies neutras a estados excitados y en última instancia a la ionización.

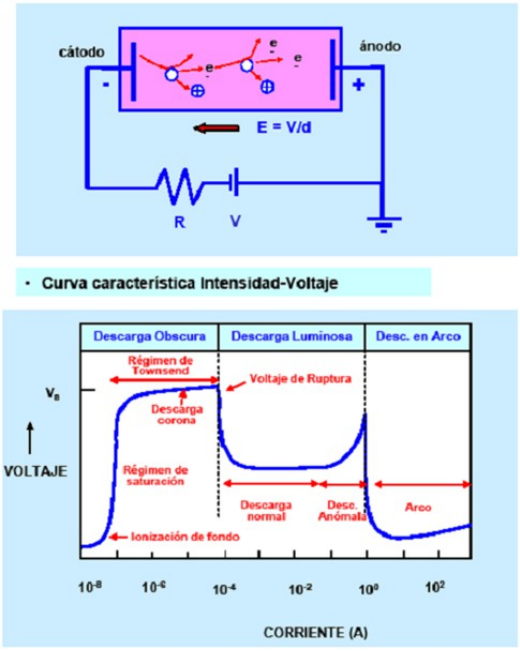


Figura 7. Esquema de una descarga.

2.9.- Ley de Paschen:
•Curva cualitativa del Voltaje de ruptura, V_B, para el H_2, N_2, Ar, Ne, aire, etc. para diferentes presiones, p, y distancia entre los electrodos de la descarga, d:

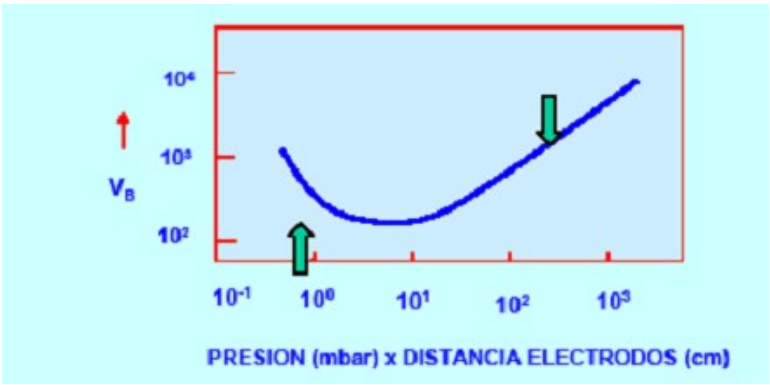


Figura 8. Curva de Paschen.

Nota: Por encima del mínimo, V_B aumenta con el valor del producto p·d debido a que el recorrido libre medio de los electrones se hace cada vez menor (al aumentar p) o a que el campo eléctrico se hace más bajo (al aumentar d). En ambos casos, la energía que adquieren los electrones entre cada colisión (como consecuencia de la aceleración producida por el campo eléctrico) es cada vez menor.
Por debajo del mínimo, el voltaje V_B crece al disminuir el valor de p·d como consecuencia de que el número de colisiones ionizantes se hace más reducido (cuando disminuye p) o bien no hay espacio suficiente para que se desarrolle la avalancha necesaria para la ruptura.

2.10.- Plasmas térmicos y plasmas fríos (no térmicos)
Plasmas fríos:
•Se producen en descargas eléctricas a presión baja cuando la corriente de descarga es pequeña. En estas condiciones el plasma está débilmente ionizado.
•Los electrones pueden ser acelerados con velocidad muy alta debido a su pequeña masa (frente a la de los iones).
•La energía cinética media de los electrones alcanza valores entre 1 y 10 eV, siendo mucho mayor que la de los iones y átomos neutros (esta última está determinada por la temperatura del medio)

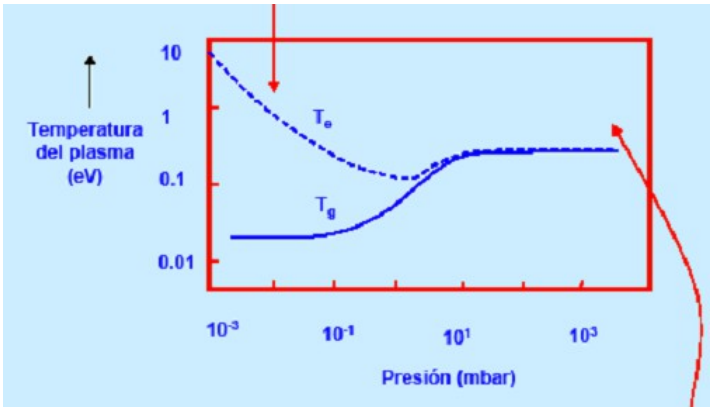


Figura 9. Plasmas fríos y térmicos

Plasmas térmicos:
•En cambio, a presiones más elevadas (p. e. presión atmosférica), la interacción entre los electrones y los átomos neutros del plasma es mucho mayor por lo que todo el conjunto alcanza la misma temperatura. Es el caso de las llamas o antorchas de plasma.

2.11.- Comparación de los parámetros característicos:
•Comparación con otros tipos de plasmas:

Parámetro (rango de valores típicos)	Plasmas fríos	Plasmas térmicos
Presión del gas (mbar)	10^{-3} - 10^{2}	10^{2} - 10^{4}
Voltaje electrodos	100 - 1000	1 - 100
Corriente (A)	0.01 - 1.0	1.0 - 100

Potencia (W)	1 - 50	$1 - 10^4$
Densidad del plasma (electrones/cm^3)	10^9 - 10^{12}	10^{14} - 10^{18}
Temperatura electrones, T_e (K)	1 - $5x10^4$	0.1 - 1
Temperatura del gas, T_g (K)	$3x10^2$ - 10^3	$T_g = T_e$

2.12.- Aplicación de los plasmas en el procesado de materiales.

Las descargas eléctricas dan lugar a la producción de numerosas especies (electrones, iones, radicales, átomos excitados) que pueden ser utilizadas en una gran variedad de procesos.

Entre ellas cabe destacar:

•Deposición de capas delgadas (técnicas de CVD y PVD).
•Ataque de superficies y capas delgadas (litografía) (Técnicas de ataque por plasma).
•Producción de haces iones para bombardeo de superficies (fuentes de iones).
•Producción de nuevas especies químicas (ozono).
•Tratamiento de superficies (funcionalización) y de eliminación de residuos.
•Producción de vacío (bombas iónicas) y de medida del vacío (manómetros de ionización).
•Depósito de materiales de alto punto de fusión (técnicas de proyección por plasma).
•Soldadura eléctrica.
•Fusión por láser.
•Lámparas de descarga (tubos de Rayos X).
•Iluminación (tubos de Ne)

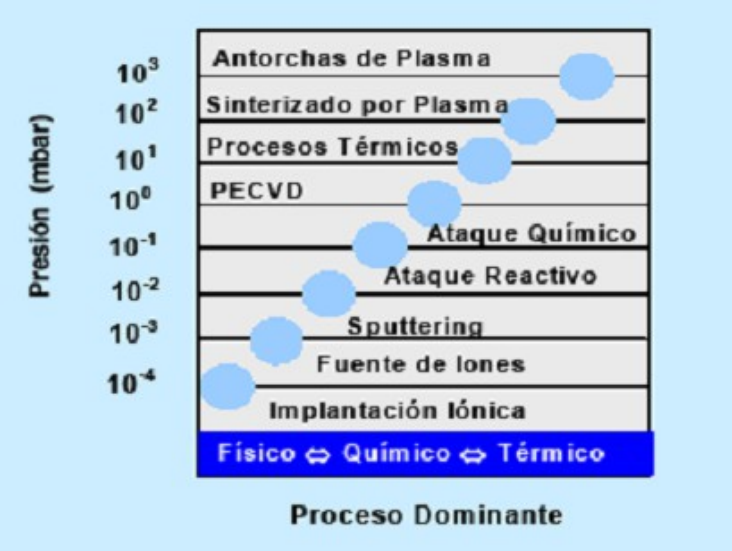


Figura 10. Procesos dominantes en función de la presión.

2.13.- Efecto de la presión.

•Presiones bajas (< 10^{-3} mbar) → predominio de fenómenos físicos (bombardeo iónico).
•Presiones intermedias (10^{-3}–10 mbar) → Se superponen fenómenos químicos (reacciones de deposición o ataque).
•Presiones altas (< 10 mbar) → Procesos térmicos (sinterización, fusión, etc.).

3. Interacción Plasma-Superficie

La energía de los electrones e iones en el plasma es suficiente para ionizar los átomos y/o moléculas neutras presentes, disgregándolos para formar iones reactivos y calentar localmente la superficie. Dependiendo del gas y de los parámetros de operación, los plasmas son capaces de realizar trabajo mecánico a través de un proceso de ablación por la transferencia cinética de electrones e iones con la superficie. También es posible realizar trabajo químico a través de la interacción de las especies iónicas reactivas con la superficie. En general, los plasmas son capaces de interactuar y modificar la superficie de los materiales a través de varios mecanismos: ablación, activación, depositación, formación de enlaces cruzados (cross-linking) e implantación (grafting).

3.1 Ablación

La ablación por plasma permite la remoción mecánica de contaminantes superficiales mediante el bombardeo de iones y electrones de alta energía. Las películas superficiales contaminantes (aceites de corte, aceites protectores, polvos desmoldadores) están unidos típicamente por enlaces débiles C-H. La ablación rompe los enlaces covalentes débiles de los contaminantes poliméricos mediante el bombardeo mecánico. Los contaminantes superficiales experimentan cortes repetitivos de sus cadenas hasta que su peso molecular es lo suficientemente bajo para ser evaporado en las condiciones de vacío. La ablación solo afecta las películas contaminantes y las capas más superficiales del material del sustrato. El gas argón es el preferido para este proceso, pues tienen una mayor eficiencia de ablación y es químicamente inerte con la superficie de los materiales.

3.2 Activación

La activación de una superficie por plasma requiere de la creación de grupos funcionales químicos sobre la superficie a través de plasma de gases como son: el oxígeno, hidrogeno, nitrógeno y amoníaco, los cuales se disocian y reaccionan con la superficie. En el caso de los polímeros, la activación de la superficie implica la sustitución de grupos poliméricos superficiales por grupos químicos provenientes del plasma. El plasma rompe los enlaces débiles de la superficie del polímero y los reemplaza con grupos altamente reactivos como son: los carbonilos, carboxilos e hidroxilos. El proceso de activación altera la actividad química y las características de la superficie, modificando la humectabilidad y la adhesión generando superficies con una alta adhesión y por lo tanto una gran duración del adhesivo.

3.3 Enlazamiento cruzado (Cross-linking):

El enlazamiento cruzado (Cross-linking) en los polímeros se refiere al establecimiento de enlaces químicos entre las cadenas moleculares de los polímeros. El plasma con gases inertes puede ser empleado para generar enlaces cruzados en la superficie del polímero y producir una película superficial más dura y resistente que la matriz. Bajo ciertas circunstancias, el enlazamiento cruzado generado por plasma puede generar una mayor resistencia química y al desgaste de la superficie polimérica.

3.4 Depositación:
Durante el proceso de depositación por plasma se forma una película delgada de polímero sobre la superficie a través del proceso de polimerización en plasma. La película delgada depositada puede presentar varias propiedades o características físicas resultantes del gas específico empleado en el plasma así como de los parámetros empleados. Los recubrimientos pueden mostrar un alto grado de entrecruzamiento y una adherencia mucho más fuerte al sustrato en comparación con películas producidas por los métodos convencionales de polimerización.

3.5 Gases empleados en limpieza y modificación superficial por plasma:

Aire
- Remoción de contaminantes (químico)
- Proceso de oxidación
- Activación de Superficies

O2
- Remoción de contaminantes (químico)
- Proceso de oxidación
- Activación de superficie (humectabilidad & adhesión)
- Ataque químico (orgánico)
- Depositación (óxidos metálicos)

Nota: debe emplearse una bomba de vacío especial para uso con oxígeno para evitar riesgos innecesarios y evitar los posibles daños.

N_2
- Activación de superficie
- Depositación (SiN (con Si), nitruros metálicos (con M))

Ar
- Remoción de contaminantes (ablación)
- Enlaces cruzados (Crosslinking)

H_2
- Remoción de contaminantes (química)
- Modificación superficial (curado)
- Procesos de reducción (óxidos metálicos)
- Depositación (metales (w/ M))

Nota: Debe tomarse extrema precaución cuando se trabaja con H_2 como gas de plasma para evitar posibles riesgos y daños.

3.6 Ventajas Generales del Plasma
El tratamiento con plasma solo afecta la región cercana a la superficie del material tratado; por lo que no modifica las propiedades de la matriz del material tratado. El proceso de limpieza superficial con plasma no deja residuos orgánicos sobre la superficie, como sucede con muchos procesos húmedos de limpieza, bajo condiciones adecuadas puede realizarse una completa remoción de los contaminantes superficiales, generando una superficie atómicamente limpia. El plasma no tiene limitaciones debidas

a efectos de tensión superficial, como las de las soluciones acuosas, por lo que puede limpiar superficies rugosas, porosas o irregulares. El tratamiento con plasma ocurre a temperaturas cercanas a la temperatura ambiental, minimizando el riesgo de dañar materiales sensibles al incremento de la temperatura.
Flexibilidad y consistencia del proceso con plasma

Dependiendo de los parámetros de operación durante el tratamiento con plasma, el proceso puede emplearse para realizar limpieza, activación, esterilización y alteración de las características superficiales de los materiales tratados. El plasma es capaz de reaccionar con una gran variedad de materiales, no solo eso, también es posible el tratamiento de dispositivos ensamblados de diferentes materiales. Es posible realizar la limpieza con plasma de partes irregulares de geometrías difíciles de limpiar mediante otras técnicas. El proceso es altamente reproducible, caracterizándose por una gran consistencia del resultado, mayor a los procesos químicos y mecánicos.

3.6.1 Bajo costo - Facilidad de empleo
El proceso con plasma es altamente eficiente con tiempos de tratamiento cortos, sin etapas de secado y consumos pequeños de energía. También reduce o incluso evita procesos adicionales de corrección de las superficies por daño térmico o por solventes. Es más fácil de usar y mantener comparado con los procesos químicos y mecánicos, además, no requiere de sistemas complicados de análisis químico o de mantenimiento. El empleo de plasma frecuentemente elimina la necesidad de solventes, junto con los costos de adquisición y disposición de los mismos.

3.6.2 Usos y Seguridad ambiental
El empleo de plasma elimina los riesgos de seguridad asociados a la exposición de operarios a químicos peligrosos. El plasma es contenido en una cámara de reacción en condiciones de vacío, con muy poca o nula exposición del trabajador. Las temperaturas de operación están cercanas a la temperatura ambiental con casi nulo riesgo de exposición a altas temperaturas. El tratamiento con plasma no0 emplea reactivos peligrosos, como son: fluorocarbonos clorinados, solventes orgánicos, químicos de limpieza ácidos. La EPA ha clasificado la mayoría de los procesos con plasma como procesos verdes y amigables con el medio ambiente

4. Aplicaciones del plasma
El proceso de plasma puede ser empleado en una gran variedad de aplicaciones, los cuales incluyen los campos de la ciencia e ingeniería de materiales (polímeros, materiales biomédicos), microfluidos, óptica, microscopia e investigación dental y médica.

4.1 Beneficios de la limpieza con Plasma:

- Remueve contaminantes orgánicos por acción química (Plasma de O_2 y de aire) o por ablación física (plasma de argón). Elimina el uso de solventes químicos, su almacenaje y la disposición de solventes gastados. Es posible limpiar superficies

con porosidad a microescala o con micro canales, superficies no deseables para limpieza con solventes debido a las limitaciones impuestas por la tensión superficial. Es capaz de volver hidrofílicas a la mayoría de las superficies, disminuyendo el ángulo de contacto del agua, incrementando la mojabilidad de la superficie tratada con plasma, ver figura 11.

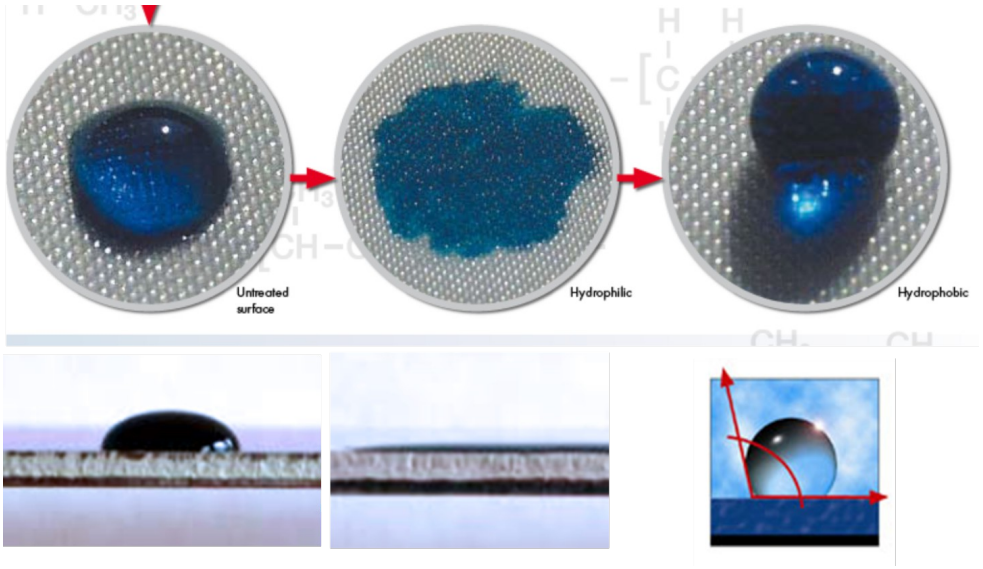


Figura 11. Se observa cómo se afecta el ángulo de mojado de la superficie como resultado del tratamiento con plasma [1, 2].

Promueve la adhesión e incrementa el pegado a otras superficies. Prepara las superficies para procesos posteriores (depositación de películas delgadas o adsorción de moléculas, proceso conocido como funcionalización de la superficie). Esteriliza y remueve contaminantes microbianos de las superficies tratadas, benéfico en aplicaciones biomédicas y biomateriales. Limpia la superficie sin afectar las propiedades de la matriz de los materiales. Puede tratarse una gran variedad de materiales aún con superficies geométricas complicadas.

Las superficies limpias disminuyen notoriamente la señal de fondo por auto-fluorescencia originada por la presencia de contaminantes orgánicos superficiales durante estudios por microscopia de fluorescencia.

4.1.1 Usos en limpieza

- Limpieza de obleas de semiconductores y sustratos (Si, Ge)
- Limpieza de diapositivas de vidrio y sus substratos
- Óxidos (Cuarzo, óxido de estaño –indio (ITO), TiO_2, Al_2O_3); mica

- Limpieza de sistemas ópticos, fibra óptica, lentes (cuarzo, Ge, ZnSe), cuvettes y sustratos empleados en mediciones espectroscópicas (ATR-FTIR, UV-Vis, SERS)
- Limpieza de cristales de cuarzo empleados en micro balanzas (QCM)
- Limpieza de las puntas de cantiléver para mediciones de morfología superficial y fuerza de fricción.
- Limpieza de rejillas empleadas como porta muestras en el análisis por microscopia electrónica (SEM y TEM)
- Limpieza de tarjetas de circuitos impresos y superficies impresas previa al pegado.
- Limpieza de superficies de oro para experimentos de auto ensamblado y de superficies metálicas en general.

Plasma de oxígeno o aire

- Remueve contaminantes orgánicos con radicales de oxigeno altamente reactivos y por ablación de iones de oxígeno altamente energéticos
- Promueve la hidroxilación de la superficie (formación de grupos OH)
- Promueve la oxidación de la superficie, la oxidación puede ser indeseable para algunas superficies metálicas (oro) lo cual afecta las propiedades superficiales.

Plasma de argón

- Limpia las superficies por el bombardeo de iones de argón y por ablación física de los contaminantes superficiales.
- No reacciona con la superficie o altera la química superficial al tratarse de un gas inerte.

En aplicaciones que son sensibles a la contaminación potencial de impurezas a nivel de trazas de elementos como el Ca, K y Na, presentes en vidrios de borosilicato, se recomienda el uso de cámaras de cuarzo en lugar de las cámaras estándar de vidrio Pyrex.

Los parámetros empleados durante el proceso de limpieza deben ser ajustados por experimentación para encontrar las condiciones óptimas de operación. Las condiciones típicas para este proceso dan a continuación:

- Presión: 100 mTorr a 1 Torr
- Potencia de RF: Medio a alto
- Tiempo de proceso: 1-3 minutos
- Potencia baja de RF: empleada para minimizar la rugosidad generada en la superficie, el tiempo de proceso debe reajustarse para compensar la baja potencia.

4.2 Beneficios del tratamiento con plasma en la fabricación de micro dispositivos:

- Tratamiento rápido de oxidación por plasma de superficies hidrofóbicas como: poli(dimetilsiloxano) (PDMS), vidrio y otros polímeros para generar superficies hidrofilicas, ver figura 12. Deben evitarse los tratamientos largos con el plasma porque pueden causar el agrietamiento del PDMS y la migración hacia la superficie de moléculas de bajo peso molecular, disminuyendo el número de grupos hidrofilicos de SiOH resultando en un pegado débil o incompleto. La superficie de PDMS recupera sus propiedades hidrofóbicas (envejecimiento) con el tiempo después del tratamiento con plasma (~1 hora).
- Pegado por oxidación de superficies de PDMS y sellado irreversible para crear canales resistentes a las fugas en dispositivos microfluidicos. Las superficies oxidadas deben ser puestas en contacto inmediatamente al tratamiento con plasma para generar el pegado más fuerte posible.
- Las superficies hidrofilicas aumentan el flujo de fluido y mojado de canales en dispositivos micro fluidicos.
- Es posible crear superficies con patrones alternantes de regiones hidrofilicas-hidrofóbicas.

4.2.1 Aplicaciones

- Estudio de reacciones químicas y flujo de fluidos a microescala.
- Detección de organismos biológicos o de especies químicas.
- Diagnóstico clínico y separación de dogas en investigación médica.
- Manipulación de fluidos a escala celular (micras) en investigación biológica.
- Crecimiento de células y cultivos de tejidos.

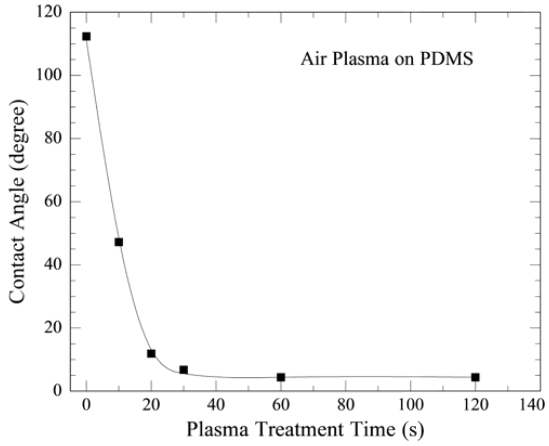


Figura 12. Ángulo de contacto de agua sobre PDMS como función del tiempo de tratamiento con plasma de aire [3, 4].

4.3 Activación y modificación de superficies:

- Modificación de superficies por adhesión o adsorción de grupos funcionales para ajustar las propiedades superficiales para aplicaciones específicas.
- Restructuración de superficies poliméricas por enlazamiento cruzado (crosslinking)
- Depósito de películas delgadas de polímeros por polimerización por plasma, ver figura 13.
- Inserción de polímeros funcionales o grupos terminales en superficies activadas por plasma, ver figura 14.
- Preparación de superficies para procesos posteriores, depositación de películas delgadas o adsorción de moléculas específicas.
- Mejoramiento de la cobertura superficial y dispersión de recubrimientos y de la adhesión entre dos superficies.
- Modificación de la mojabilidad de superficies hidrofóbicas e hidrofilicas , ver figuras 15 y 16, con el gas o gases apropiados de proceso de plasma.
- Cambio de las propiedades superficiales sin afectar las propiedades de la matriz de los materiales.

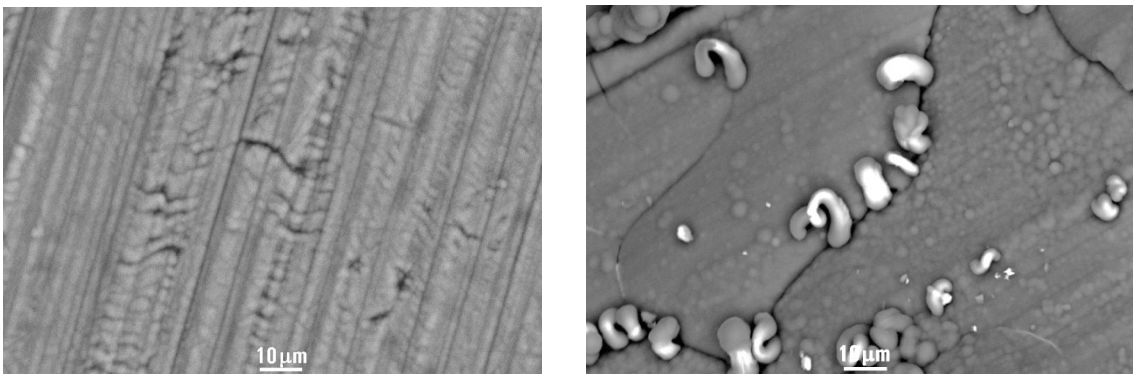


Figura 13. Deposito a partir de diclorometano sobre sustrato de cobre, (a) Cobre puro, (b) deposito a partir de diclorometano.

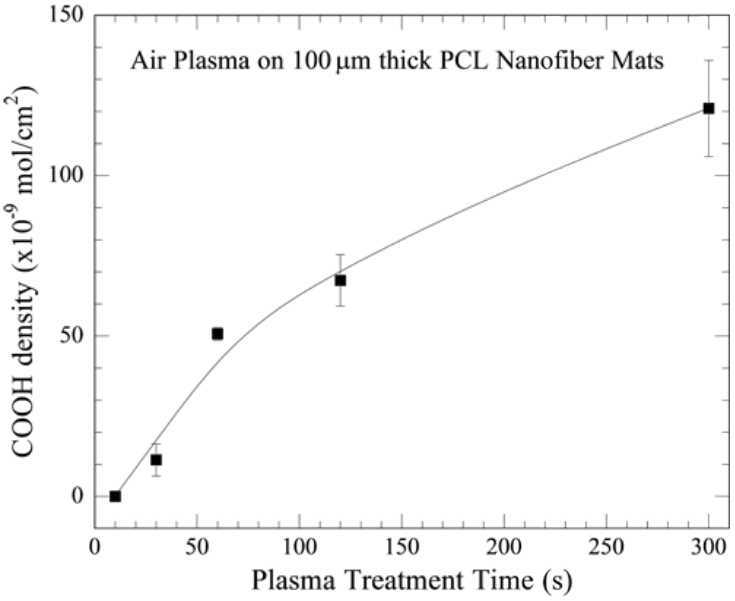


Figura 14. Densidad superficial de grupos carboxílicos (COOH) como función del tiempo de tratamiento con plasma de aire sobre una superficie de 100 μm de espesor de tejidos de nanofibras de poli(caprolactona) (PCL). Las capas de COOH facilitan la inserción subsecuente de moléculas de gelatina dentro de las fibras del tejido de PCL con uso potencial como soportes en ingeniería de tejidos [5].

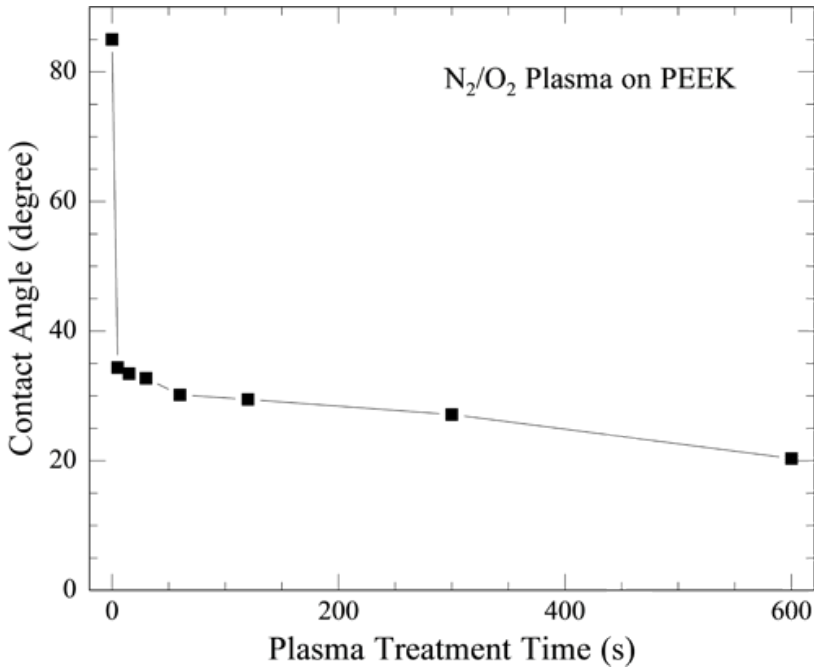


Figura 15. Ángulo de contacto como función del tiempo de tratamiento con plasma de N_2/O_2 sobre polieter-eterketona (PEEK). La superficie de PEEK se vuelve hidrofilica después de 20 segundos de tratamiento con plasma [6].

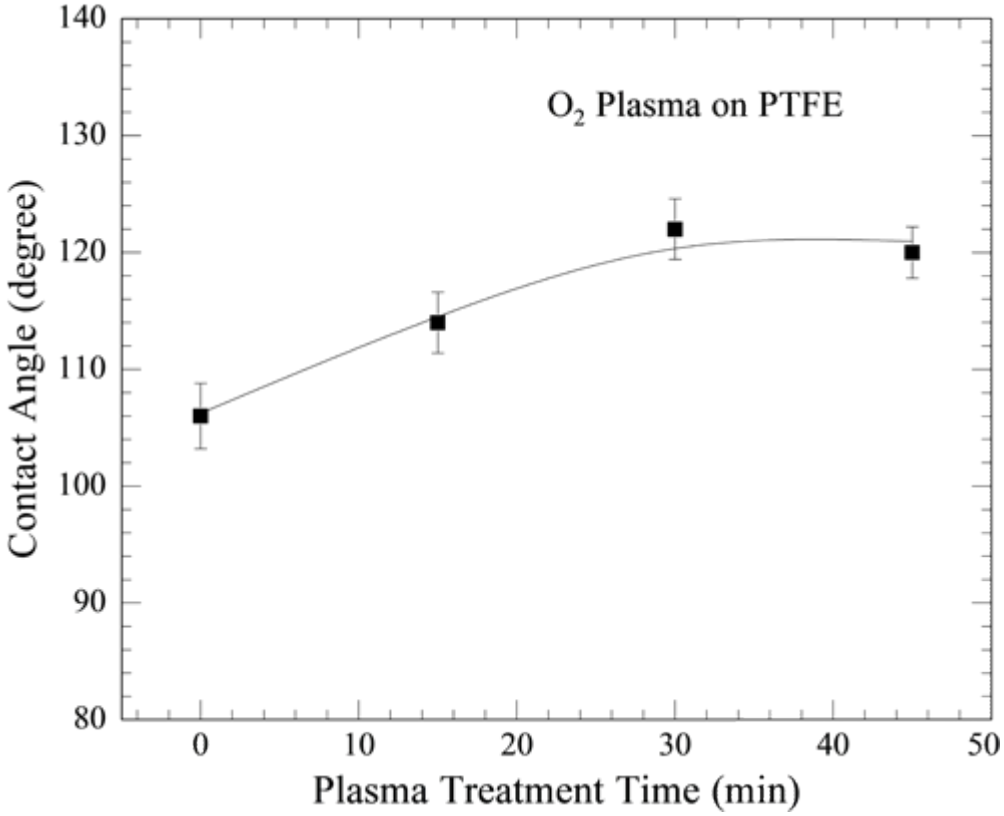


Figura 16. Ángulo de contacto como función del tiempo de tratamiento con un plasma de aire sobre poli(tetrafluoroetileno) (PTFE), mostrando un incremento en la hidrofobicidad de la superficie. El tratamiento con plasma genera rugosidad a nanoescala que también incrementa la hidrofobicidad de la superficie [7].

4.3.1 Aplicaciones:

- Estudios de auto-ensamble con superficies con patrones de regiones hidrofóbicas e hidrofilicas.
- Investigación dental de adhesión de células periodontales.
- Promoción de la adhesión de microorganismos sobre superficies tratadas con plasma.
- Promoción de adhesión de células y proliferación de células sobre biomateriales y soportes de tejidos modificados con plasma.
- Modificación de superficies para actuar como capas protectoras o barreras a la matriz del material.
- Generación de superficies poliméricas con enlazamiento cruzado para reducir la permeabilidad a moléculas específicas.
- Proporcionar por oxidación y por la formación de grupos hidroxilos (OH) hidrofilidad a superficies.
- Proporcionar hidrofobicidad a superficies por la depositación de grupos que contienen fluoruros (CF, CF_2, CF_3)

4.3.2 Métodos de procesamiento para plasma de tetrafluoruro de carbono (CF_4)

    - Forma recubrimientos hidrofóbicas de grupos fluorinados (CF, CF_2, CF_3)
    - Disminuye el número de grupos polares hidrofilicos sobre la superficie, disminuyendo la humectabilidad de la superficie.
- Las superficies deben ser empleadas inmediatamente después del tratamiento con plasma, puesto que estas recuperan sus características superficiales al estar expuestas al aire por tiempos prolongados.
- Los parámetros sugeridos de proceso para el tratamiento con plasma de CF_4 se indican a continuación aun cuando estos deben ser ajustados experimentalmente para lograr las condiciones óptimas de proceso.
    - Presión: 100 mTorr a 1 Torr
    - Potencia de RF: Medio a alta
    - Tiempo de proceso: 1-3 minutos

4.4 Adhesión y humectabilidad superficial

- El tratamiento con plasma de fibras mejora la adhesión a la matriz en materiales compuestos con fibras de refuerzo
- Permite el estudio de las características de adhesión de materiales disimiles durante pruebas mecánicas o por mediciones de esfuerzo mediante microscopia de fuerza atómica (AFM).

4.4.1. Biomateriales

- Incrementa la adhesión y modifica las propiedades de mojado de las superficies de biomateriales
    - Los biomateriales son típicamente inertes químicamente y tienen bajas energías de superficie para minimizar la contaminación por microrganismos y las interacciones indeseables con otras superficies, estas propiedades también dificultan la aplicación de recubrimientos funcionales o la adhesión de grupos molecularmente activos sobre la superficie.
    - El tratamiento con plasma puede mejorar la funcionalidad y biocompatibilidad de las superficies de los biomateriales.
- Esterilización
    - El tratamiento con plasma de oxígeno puede de manera simultánea limpiar y esterilizar la superficie de dispositivos médicos y biomateriales
    - La esterilización con plasma es apropiado para implantes médicos y dentales que son sensibles a las temperaturas elevadas, químicos o medios irradiados en autoclaves, óxido de etileno (EtO) o esterilización por radiación gamma.

4.4.1.1 Aplicaciones

- La activación por plasma de las superficies de sustratos proporciona superficies hidrofilicas y promueve el pegado y adhesión de especies biológicas funcionales o de recubrimientos.
    - Incrementa la adhesión celular, cobertura y proliferación sobre soporte de tejidos.
    - Promueve la adsorción de especies biológicas funcionales selectas mientras que inhibe la adhesión de bacterias y de microrganismos contaminantes.
    - Los recubrimientos sobre superficies de biomateriales tratados con plasma pueden actuar como barreras protectoras o lubricantes en dispositivos de implante médicos.
- El tratamiento con plasma puede limpiar y activar arreglos de microelectrodos para biosensores.
- Es posible la esterilización de dispositivos médicos y de biomateriales (implantes dentales, materiales para moldes de impresiones dentales, soportes de tejidos).

5. Agradecimientos.

Se agradece el apoyo técnico del Sr. Anselmo González y al Q. Ivan Puente, así como el apoyo financiero de los proyectos DGAPA IN-105010, IN-109511 y CONACyT 128714.

6. Referencias

1.- Plasma Surface Technology. Diener electronic GmbH + Co. KG, 2009. Internet: www.plasma.de

2.- S. Yamasaki1, H. Miyahara, R. Sasaki, R. Shimada, E. Hotta, A. Okino. High-speed cleaning/treatment of metal surface using an atmospheric damage-free plasma source. 29th ICPIG, July 12-17, 2009, Cancún, México

3.-. Jiang, X., H. Zheng, S. Gourdin, P. T. Hammond. "Polymer-on-Polymer Stamping: Universal Approaches to Chemically Patterned Surfaces." Langmuir (2002) 18: 2607-2615

4.- Zheng, H., M. F. Rubner, P. T. Hammond. "Particle Assembly on Patterned "Plus/Minus" Polyelectrolyte Surfaces Via Polymer-On-Polymer Stamping." Langmuir (2002) 18: 4505-4510.

5.- Ma, Z., W. He, T. Yong, S. Ramakrishna. "Grafting of Gelatin on Electrospun Poly(caprolactone) Nanofibers to Improve Endothelial Cell Spreading and Proliferation and to Control Cell Orientation." Tissue Eng. (2005) 11: 1149-1158.

6.- Ha, S. W., M. Kirch, F. Birchler, K.-L. Eckert, J. Mayer, E. Wintermantel, C. Sittig, I . Pfund-Klingenfuss, M. Textor, N. D. Spencer, M. Guecheva, H. Vonmont. "Surface Activation of Polyetheretherketone (PEEK) and Formation of Calcium Phosphate Coatings by Precipitation." J. Mater. Sci. Mater. Med. (1997) 8: 683-690.

Medidas de complejidad de series de tiempo fisiológicas

Ruben Fossion [1]

Instituto de Geriatría, Periférico Sur No. 2767, Col. San Jerónimo Lídice, Del. Magdalena Contreras, 10200 México D.F., México
Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 México, D.F., México

Keywords: Time-Series Analysis, Fluctuation Phenomena, Random Processes and Brownian Motion, Noise, Nonlinear Dynamics and Chaos
PACS: 05.45.Tp, 05.40.-a,02.50.-r, 05.40.Ca, 05.45.-a

Introduction

Transiciónes demográfica y epidemiológica

Las poblaciones de muchos países en el mundo se están envejeciendo a un ritmo acelerado. El drástico cambio en la estructura de tales poblaciones, con un número mayor de personas mayores que necesitan apoyarse en un número reducido de jóvenes y adultos activos, causa una gran variedad de fricciones y problemas de índole social, económico y médico. Uno de estos problemas es una transición de predominancia de enfermedades agudas y infecciosas (p.ej. gripa) a enfermedades crónicas no-transmisibles (p.ej. enfermedad de Parkinson o Alzheimer), un fenomeno descrito por primera vez por Omran en 1971 [1]. Seguridad social y el sector salud deben adaptarse a esta transición centrándose más en prevención y atención integral durante mucho tiempo en lugar de intervenciones agudas [2].

Enfermedades dinámicas y pensamiento sistémica

El mundo médico comienza a distinguir entre enfermedades “simples” y “complejas”, más allá del paradigma actual. En una enfermedad “simple”, hay un solo problema que puede resolverse con una sola intervención, y el efecto de la intervención es proporcional a la cantidad de medicamentos administrados (lineal). En una enfermedad “compleja”, hay múltiples complicaciones y la enfermedad no puede ser curada al intervenir individualmente en cada uno de los problemas, tampoco es el efecto proporcional a la cantidad de medicación administrada (no lineal) [3]. Una apendicitis puede servir como un ejemplo de una enfermedad "simple" y su operación como una cura “simple”, el

[1] fossion@nucleares.unam.mx

paciente se presenta a la sala de emergencia con dolor abdominal, los clínicos diagnostican la enfermedad y los cirujanos dan solución al problema eliminando el apéndice con un procedimiento quirúrgico. Por el contrario, el envejecimiento y sus problemas relacionados con la salud, así como otras enfermedades (fragilidad relacionada al envejecimiento, fibromialgia, trastorno funcional digestivo, síndrome de la guerra del golfo), son "complejas" dando lugar a problemas que afectan de muchas maneras a los sujetos y requiere de un enfoque "complejo", multidimensional o integral; tanto para su abordaje como para su tratamiento. Glass, Mackey et al. explican los trastornos "complejos" como *enfermedades dinámicas*, donde hay alteraciones en la sincronización de ritmos y en los ritmos mismos de diferentes procesos biológicos del organismo [4, 5, 6, 7]. Estas enfermedades son dificilmente detectable o tratable en el paradigma actual de la Medicina occidental que es visual y reduccionista. En cambio, Ahn et al. abogan por un *pensamiento sistémico*, donde más que enfocarse en un sólo componente, se interpreta el organismo como una red de órganos y procesos en interacción [8, 9]. En contraste, para la Medicina Tradicional China (TCM por sus siglas en inglés), que es holística, la mayoría de las enfermedades son dinámicas [10, 11].

De la homeostasis a la homeodinámica o la homeocinética

El propósito del organimso como sistema es crear un *medio interno* que puede funcionar independiente de las condiciones variables (temperatura, humedad, etc.) del *medio externo* del ambiente. Las plantas, los vertebrados de sangre fría a y los vertebrados homotermos son sucesivamente más exitosos en independizar su medio interno de su medio ambiente. Primero se pensó que este medio interno fuera estático (Claude Bernard), luego se incluyó el concepto de que el medio interno después de perturbaciones regresa a su "steady state", un mecanismo que se conoce como *homeostasis* (Walter B. Cannon). En cambio, resulta que el medio interno logra su estabilidad de una manera dinámica, apoyándose en una red de mecanismos de retroalimentación positiva y negativa (Lawrence J. Henderson, Ludwig von Bertalanffy, Norbert Wiener), un fenómeno que algunos prefieren llamar *homeodynámica* o *homeocinética* [12, 13]. Goldberger y colaboradores encontraron que cuando se estudia la *serie de tiempo* de cualquier observable fisiológico, es decir, la manera en la cual evoluciona en el tiempo, se observa que fluctúa de manera continua alrededor de un promedio adecuado [14, 15]. Condiciones cambiables del medio ambiente al cual el medio interno debe de adaptarse constituyen *estresores*, y la adaptabilidad del organimso depende de la fuerza de la respuesta al estresor, véase la Fig. 1. La adaptabilidad resulta suboptimal cuando la respuesta es deficiente o excesiva, lo cual se llama allostasis o cacostasis, y lo cual que es nocivo para la salud. La adaptabilidad es optimal cuando la fuerza de respuesta se limita a cierto rango adecuado. A este estado resultante se sigue llamando – por razones históricos – homeostasis. De esta manera, Chrousos logra explicar una gran variedad de enfermedades como síntomas diferentes del mismo síndrome de un sistema de estrés mal adaptable, véase la Tabla 1.

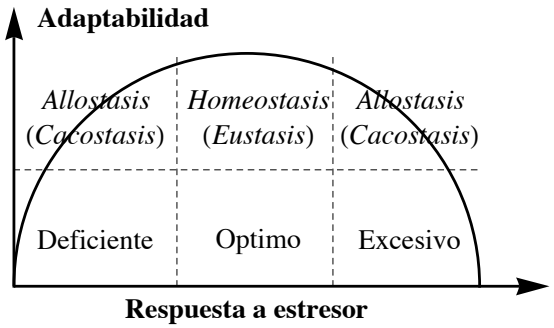


FIGURE 1. Esquema de la adaptabilidad de un organismo en función de la fuerza de su respuesta a estresores interiores o exteriores. Cuando la respuesta es deficiente o excesiva, la adaptabilidad es suboptimal (allostasis o cacostasis), lo cual puede ser nocivo a corto o largo plazo. Sólo cuando la respuesta es adecuada (homeostasis o eustasis), la adaptabilidad es optimal. En este esquema, se pueden explicar una gran variedad de enfermedades como los síntomas diferentes dentro del mismo síndrome de un sistema de estrés defectuoso, véase la Tabla 1. Figura adaptada de [16, 17].

TABLE 1. Una gran variedad de enfermedades se pueden explicar como síntomas diferentes dentro del mismo síndrome de un sistema de estrés defectuoso, con respuestas deficientes o excesivas a estresores, lo cual resulta en un medio interno mal-equilibrado (allostasis o cacostasis), véase la Fig. 1. Tabla adaptada de [16, 17, 18].

Respuesta deficiencia	**Respuesta excesiva**
Insuficiencia adrenal	Síndrome de Cushing
Depresión atípica/estacional	Estrés crónico
Síndrome de fatiga crónica	Depresión melancólica
Fibromialgia	Anorexia nerviosa
Síndrome de tensión premenstrual	Trastorno obsesivo-compulsivo
Depresión climactérica	Trastorno de pánico
Abstinencia de nicotina	Alcoholismo crónico activo
Al interrumpir terapia con glucocorticoides	Ejercicio excesivo (atletismo compulsivo)
Al interrumpir tratamiento para el Sx. de Cushing	Abstinencia de alcohol y de nicotina
Después de estrés crónico	Diabetes mellitus
Período post-parto	Obesidad central (síndrome metabólico)
Síndrome de estrés postraumático en adultos	Síndrome de estrés postraumático en niños
Hipotiroidismo	Hipertiroidismo
Artritis reumatoide	Embarazo
Asma y eccema	

Fluctuaciones en series de tiempo y sus medidas

En este contexto de pensamiento sistémica de la fisiología, se puede apreciar un acercamiento entre la Medicina, las Ciencias Exactas y la Ingeniería, y más en particular en la identificación de nuevos *biomarcadores no-sintomáticos*. A diferencia del diagnóstico médico clásico, que se basa en síntomas y pérdida de funcionalidad del paciente, biomarcadores no-sintomáticos o no-funcionales muchas veces se basan en un análisis estadístico cuidadoso de señales fisiológicas. Esta nueva corriente en la Medicina viene impulsada por el desarrollo de nuevas tecnologías que son capaces de medir con una

precisión sin precedentes una gran variedad de observables fisiológicos y cómo estos fluctúan en el tiempo. Sin duda, uno de los biomarcadores no-sintomáticos más exitosos es la *Variabilidad del Ritmo Cardíaco* (HRV, por sus siglas en inglés), lo cual predice morbilidad y mortalidad por una gran variedad de enfermedades, antes de la aparición de síntomas en el paciente, basándose solamente en cómo varía el ritmo cardíaco alrededor de su promedio [19]. El razonamiento es que si un sistema produce una señal o una *serie de tiempo*, es decir, un observable del cual se puede estudiar su evolución en el tiempo, esta señal da información sobre la dinámica interna del sistema, sin necesidad de manipular el sistema o abrirlo, lo cual en la Medicina se llama un método *no-invasivo*. Si el observable es constante en el tiempo y por consecuencia la señal no fluctúa, el sistema debe de ser *estático*. Si la señal fluctúa, el sistema debe de ser *dinámico*, y la manera en la cual fluctúa delata el tipo de dinámica interna. Es una técnica que se aplica en la óptica para averiguar si una fuente de luz es térmica, coherente o cuántica [20], en la física nuclear para estudiar el fenómeno de caos cuántico [21], y más en general en el marco de las *señales de alerta temprana*, si un sistema físico, químico, económico, ecológico, biológico o médico está en una transición de fase, cerca de su punto crítico al borde de un colapso [22, 23].

Ahora que se estableció que las series de tiempo fisiológicas fluctúan contínuamente, la pregunta es cómo fluctúan y cuál es la herramienta matemática más adecuada para describir y explicar estas fluctuaciones. En principio, las fluctuaciones de un sistema dinámico pueden ser (a) complejas, (b) aleatorias, (c) caóticas, o (d) periódicas, véase la Fig. 2. El *caos* implica determinismo, es decir, que existe la posibilidad de describir la dinámica con una o más ecuaciones. Se utiliza la palabra *complejidad* para un comportamiento irregular que sin embargo contiene ciertos patrones de orden, posiblemente de origen probabilístico (es decir, no determinístico). En cuanto a la HRV, en la literatura científica médica hay decenas de índices diferentes para medir las fluctuaciones de series de tiempo cardíacas: en el dominio del tiempo, de la frecuencia, de la complejidad y de la fractalidad [24], sin embargo, en la práctica se suele utilizar un solo índice, y más en particular el índice más sencillo en el dominio del tiempo, lo cual es la desviación estándar $\sigma = \sum_{i=1}^{N} (\Delta t(i) - \langle \Delta t \rangle)/N$. Aquí, en este contexto, σ mide promediado sobre N intervalos entre latidos cardíacos sucesivos $\Delta t(i)$ la desviación del comportamiento promedio $\langle \Delta t \rangle$ [25]. En la Fig. 2, se puede apreciar que σ se maximiza para series periódicas (compáranse los paneles (a-c) con (d)). Visto que series fisiológicas sanas son muy irregulares, como en el caso de la serie compleja del panel (a), Lipsitz advierte que la "variabilidad" de la serie de tiempo, cuantificada por σ, no refleja la complejidad de los patrones escondidos en series fisiológicas sanas [26]. Además, a parte de la manera en la cual fluctúa la serie de tiempo, es decir, el orden en el cual parecen las fluctuaciones, hay otra contribución importante y diferente a σ, la cual es la amplitud de la serie de tiempo (compáranse los paneles (d) con (e)). Se ha encontrado que en caso de enfermedad, en adultos se altera sobretodo el orden de las fluctuaciones, mientras que en neonatos se observan más bien cambios en la amplitud de la serie de tiempo [27].

Desde el inicio fue claro que las fluctuaciones fisiológicas fueron ni aleatorias ni periódicas. Primero se pensó que estas flucuaciones se dejarían explicar dentro el marco de la *teoría del caos* [28]. Sin embargo, luego surgieron dudas si la dinámica de un

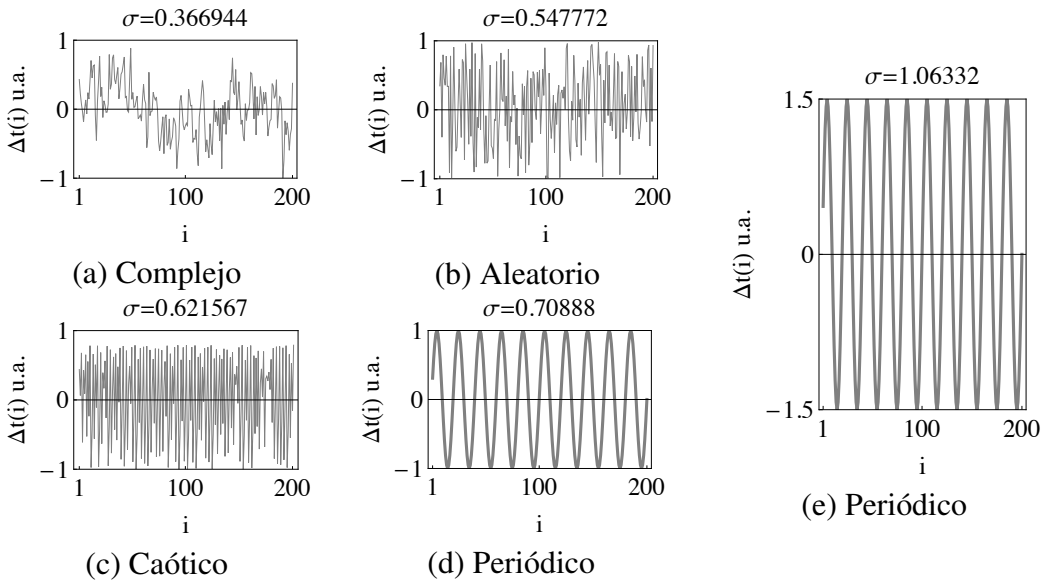


FIGURE 2. Series de tiempo esquemáticas que representan los intervalos entre latidos cardíacos sucesivos $\Delta t(i)$ en unidades arbitrarias (u.a.). (a) Serie compleja, (b) aleatoria, (c) caótica, (d) periódica y (e) periódica pero de mayor amplitud. Está también indicada la desviación estándar σ, que cuantifica la *variabilidad* de la serie de tiempo, la cual se maximiza para series periódicas, así que "variabilidad" no es sinónimo con "complejidad". Influyen en la variabilidad la manera de fluctuar, es decir, el orden en el cual parecen las fluctuaciones (compáranse los paneles (a-d)), pero también la amplitud de la serie de tiempo (compáranse los paneles (d) y (e)). Figura parcialmente inspirada en [26].

corazón sano es caótico [29], y recientemente se dedicó todo un número de la revista Chaos al tema [30]. Posiblemente los *fractales* y el nuevo marco teórico de la *complejidad* permiten describir la dinámica de las fluctuaciones fisiológicas. Se descubrió que el ritmo cardíaco tiene mucha *memoria*, es decir, que hay *correlaciones de largo alcance* entre el latido presente y muchos latidos anteriores y futuros, que se extienden hasta 24 horas o más [31, 32], un comportamiento que muchos – si no todos – los ritmos fisiológicos tienen en común [26]. Prigogine [33] y Gell-Mann [34] confirman que son las correlaciones de largo alcance en una serie de tiempo que constituyen su complejidad. En cambio, una de las características del caos es el *fenómeno mariposa*, es decir, la extrema dependencia del sistema de sus condiciones iniciales para su futura evolución en el tiempo, lo cual hace que series de tiempo caóticas carecen de memoria [35, 36]. A parte de correlaciones de larga alcance, las fluctuaciones de series de tiempo fisiológicas se comportan aproximadamente como fractales, es decir, fluctuaciones de diferentes escalas en el tiempo son similares entre si (auto-similitud) [14, 37]. Matemáticamente, el espectro de potencias de una serie fisiológica se puede aproximar con una ley de potencias,

$$P(f) \sim f^{\beta}, \tag{1}$$

con $-2 \leq \beta \leq 0$. El ruido blanco ($\beta = 0$) corresponde con una serie aleatoria como se observa en ciertas enfermedades (p.ej. fibrilación en el caso del corazón) [38, 39]. El ruido browniano ($\beta = -2$) corresponde con ruido blanco sumado, y se observa en los ritmos más regulares y rígidos de un órgano envejecido [40]. El ruido $1/f$ ($\beta = -1$)

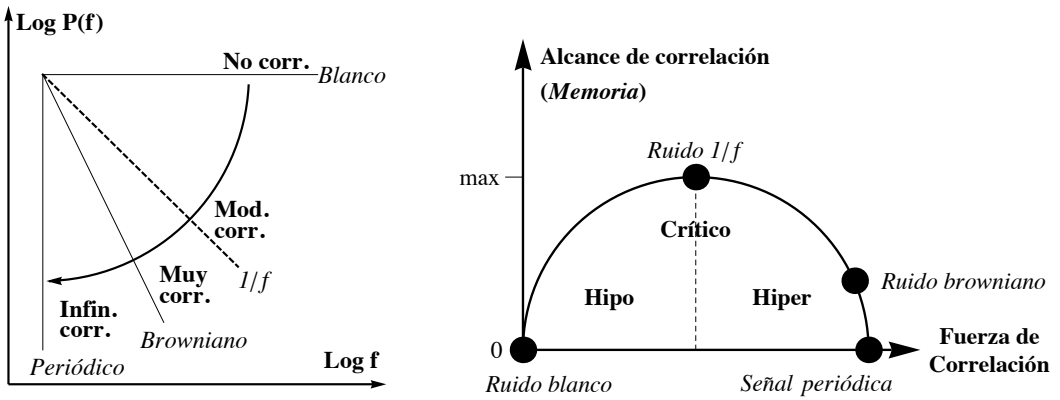
corresponde con un ritmo fisiológico sano y complejo [31, 32]. En la Fig. 3(a), se presenta un espectro de potencias esquemático en escala logarítmica. La *fuerza de la correlación* se puede estimar con,

$$|\beta| \approx \frac{|\Delta \log P|}{|\Delta \log f|}, \tag{2}$$

es decir, la fuerza de correlación es mayor mientras más potencia $\Delta \log P$ se concentra en un rango menor de frecuencias $\Delta \log f$. Así se entiende que la fuerza de correlación es nula para el ruido blanco ($\beta = 0$), moderada para el ruido $1/f$, fuerte para el ruido browniano ($\beta = -2$), e infinita para una señal periódica ($\beta = -\infty$). Lipsitz y Goldberger propusieron la *hipótesis de pérdida de complejidad con enfermedades y envejecimiento*, así que la complejidad reflejería el estado de salud del órgano que produce la serie [41]. En la Fig. 3(b) se presenta un diagrama de fase esquemático de la complejidad de las series de tiempo de tipo de la ec. (1), donde la complejidad es maximal para el ruido $1/f$, lo cual es el ruido que se observa empíricamente para series de tiempo fisiológicas sanas, y donde se pierde complejidad cuando la serie se degenera hacia el ruido blanco (perdiendo fuerza de correlación, estado hipo) o cuando la serie se degenera hacia el ruido browniano o una señal periódica (ganando fuerza de correlación, estado hiper). Nótese la similitud con la Fig. 1 del sistema homeodinámico en función de la fuerza de respuesta a estresores. En la secciones siguientes averiguaremos la validez del diagrama de fase de la Fig. 3(b) con datos experimentales y con una simulación teórica.

Pérdida o incluso aumento de la complejidad?

Un problema en el estudio de series de tiempo es la existencia de muchos índices y métodos de análisis alternativos, como p.ej. la variabilidad σ, la dimensión fractal, el análisis espectral, la entropía aproximada ApEn, el Detrended Fluctuation Analysis (DFA), etc., donde no siempre cada índice mide o cuantifica lo mismo. Probablemente, la discusión en la literatura sobre la validez de la hipótesis de Lipsitz y Goldberger sobre la pérdida de complejidad con enfermedades y el envejecimiento se debe a un malentendimiento de este tipo. Vaillancourt y Newell sostienen que la complejidad también puede aumentar en condiciones adversas de salud [44, 45]. Uno de sus argumentos se basa en la Fig. 4. La figura muestra que una pérdida de complejidad en series de tiempo se debe a la pérdida de componentes o interacciones en el sistema dinámico que genera la serie, pero sugiere que la aparición de nuevos componentes y interacciones causará un aumento en la complejidad de la serie de tiempo. Goldberger, Lipsitz et al. respondieron que Vaillancourt y Newell se basaban en el índice de la entropía aproximada ApEn, lo cual mide la impredecibilidad de la serie de tiempo. La impredecibilidad de la serie de tiempo aumenta cuando la serie de tiempo se degenera hacia el ruido blanco, véase la Fig. 13. Esto no corresponde con un aumento de la complejidad pero más bien con una pérdida [46]. La comprobación del diagrama de fase de la Fig. 3 demostraría la hipótesis de Lipsitz y Goldberger que la complejidad se maximiza en el ruido $1/f$ y que se pierde con enfermedades y envejecimiento. Más que una discusión académica sobre análisis



(a) Espectro de potencias esquemático

(b) Diagrama de fase esquemático

FIGURE 3. (a) Espectro de potencias esquemático en escala logarítmica, $P(f) \sim f^{\beta}$ ($-\infty \leq \beta \leq 0$). Un extremo es el espectro plano del ruido blanco ($\beta = 0$), donde no hay correlaciones pero al cual contribuyen muchas frecuencias. Otro extremo es el espectro vertical de una señal periódica ($\beta = -\infty$), que está maximalmente correlacionada pero al cual contribuye una sola frecuencia. El ruido $1/f$ ($\beta = -1$) está en el justo medio entre los dos extremos. (b) Diagrama de fase esquemático para los ruidos fractales, donde se muestra el comportamiento de la complejidad (la memoria o el alcance de las correlaciones) en función de la fuerza de correlación β. La complejidad se maximiza en el ruido $1/f$ (estado crítico), y se pierde complejidad cuando la fuerza de correlación disminuye hacia el ruido blanco (estado hipo) y también cuando la fuerza de correlación aumenta hacia una señal periódica (estado híper). Nótese la similitud con el esquema de Chrousos de la adaptabilidad del sistema homeodinámico de la Fig. 1. Figuras (a) y (b) adaptadas de [37, 42, 43]

estadísticos alternativos, lo anterior refleja la búsqueda del biomarcador no-sintomático mejor en series de tiempo. Un índice como la memoria o el alcance de las correlaciones que se maximiza en el estado de salud y que se pierde progresivamente en condiciones adversas, puede servir como tal biomarcador. Un índice como la ApEn que no tiene un comportamiento particular en el estado de salud y que se maximiza para el ruido blanco no sirve para este propósito.

Datos fisiológicos

El propósito de esta sección es verificar la hipótesis del diagrama de fase de la complejidad en función de la fuerza de la correlación (Fig. 3) con datos fisiológicos experimentales. Se presentan datos de (a) un sujeto con enfermedad coronaria, (b) un sujeto sano, y (c) un sujeto con muerte cerebral, véase la Fig. 5. Los datos consisten en series de 1000 intervalos Δt entre latidos sucesivos en el corazón, medidos en milisegundos. Los datos se adquirieron con el dispositivo comercial Polar [47]. Los datos cardíacos[2] son una cortesía del Dr. M.F. de Godoy, de la Faculdade de Medicina da

[2] Los datos originales incluyeron 4 series de jóvenes sanos, 3 casos de muerte cerebral y 3 pacientes con enfermedad cardíaca coronaria. Los exponentes espectrales que resultan del análisis son $\beta_{\text{sano}} =$

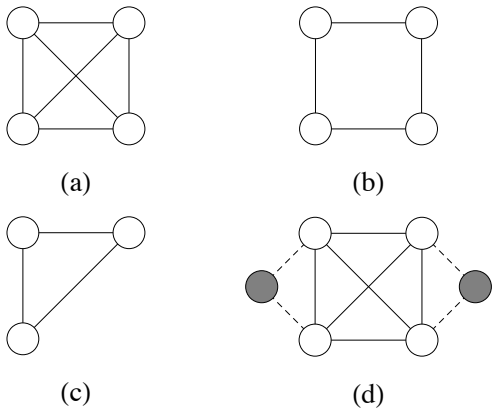


FIGURE 4. Organismo que consiste de un sistema de componentes (discos) en interacción (lineas). (a) Organismo sano, (b) pérdida de complejidad por pérdida de interacciones, (c) pérdida de complejidad por pérdida de un componente, (d) supuesto aumento de la complejidad por la aparición de nuevos componentes (p.ej. cáncer). Figura adaptada de [44]

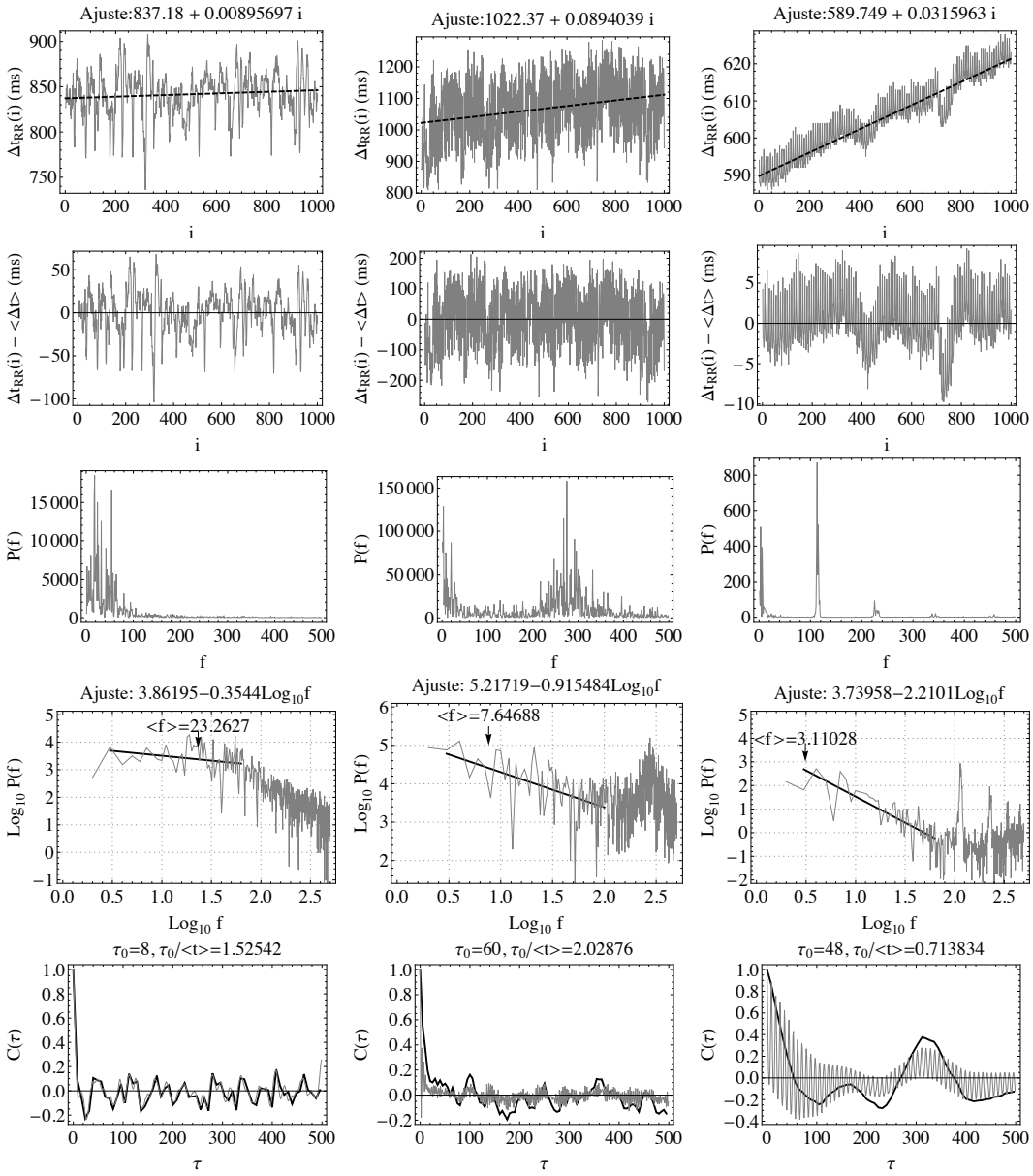
Universidade de São José do Rio Preto (FAMERP), São Paulo, Brasil [48]. Los cálculos se hicieron con Wolfram Mathematica versión 8.0

En la Fig. 5, en la primera fila, se presentan las tres series de tiempo. Nótese que la amplitud de la serie de tiempo es mayor para el sujeto sano, es más pequeña en el caso de la enfermedad cardíaca coronaria y es mínima en el caso de la muerte cerebral. Las series fluctúan alrededor de un promedio $\langle \Delta t \rangle$ que aumenta linealmente en el tiempo.

En la segunda fila, se muestran las series de fluctuaciones, $\Delta t(i) - \langle \Delta t \rangle$ alrededor del comportamiento promedio.

En la tercera fila, se dan los espectros de potencia $P(f)$, que se calcularon tomando el cuadrado de la transformada de Fourier (FFT) de la serie de fluctuaciones. El espectro de potencias completo consiste de dos mitades simétricas (una relacionada con los números reales y otra con los números imaginarios), se presenta la primera mitad del espectro de potencias. La frecuencia f expresa cuántas veces ocurre un evento en toda la duración t_{ancho} de la serie de tiempo, donde $t_{\text{ancho}} = \sum_{i=1}^{N} \Delta t(i) \approx N \langle \Delta t \rangle \approx N t_{\text{pix}}$, y donde $N = 1000$, $\langle \Delta t \rangle$ es el intervalo promedio de la serie de tiempo y $t_{\text{pix}} \approx \langle \Delta t \rangle$ es la precisión o "pixel" en el tiempo. En la Tabla 2, se traduce la frecuencia f de ocurrencias por duración de tiempo a las unidades Hz y latidos por minuto (bpm, por sus siglas en inglés). En el caso sano y el caso de la muerte cerebral, el pico entre las 200 y 300 ocurrencias por t_{ancho} corresponde con el ritmo de la respiración que influye directamente en el ritmo cardíaco. En el caso sano, este ritmo se traduce a una respiración cada 3.5 a 5 segundos, en el caso de muerte cerebral, hay una respiración cada 2 a 3 segundos. Se considera normal el ritmo de una respiracón cada 3 segundos. En el caso

$-0.92, -0.19, -0.57, -0.57$, $\beta_{\text{muerte}} = -2.2, -0.78, -1.86$ y $\beta_{\text{coronaria}} = -0.74, -1.20, -1.21$. Seleccionamos de cada categoría la serie que mejor corresponde con el modelo de Goldberger y Lipsitz.



(a) Enfermedad coronaria (b) Sano (c) Muerte cerebral

FIGURE 5. Series de tiempo y algunos análisis estadísticos para (a) un sujeto con enfermedad cardíaca coronaria, (b) un sujeto sano, y (c) un sujeto con muerte cerebral. Presentados son, la serie de tiempo de intervalos entre latidos cardíacos sucesivos $\Delta t(i)$ con un ajuste lineal del promedio (primera fila), las fluctuaciones de la serie de tiempo $\Delta t(i) - \langle \Delta t \rangle$ (segunda fila), espectro de potencias $P(f)$ de las fluctuaciones (tercera fila), espectro de potencias en escala logarítmica con ajuste de la parte que sigue una ley de potencias y con la frecuencia promedia $\langle f \rangle$ (cuarta fila), función de autocorrelación $C(\tau)$ de la serie completa (gris) y de la parte de la serie que corresponde con la ley de potencias (negro) (última fila). La complejidad de la serie de tiempo se puede cuantificar con su *memoria* (el tiempo que necesita la función de autocorrelación para decaer hasta 0). Se puede observar que la memoria se maximiza para el sujeto sano (τ_0 en tiempo absoluto y $\tau_0/\langle t \rangle$ en tiempo propio de la serie de tiempo). Los datos cardíacos son una cortesía del Dr. de Godoy [48].

del sujeto sano y de la enfermedad coronaria se trata de espectros de banda ancha, que se extienden a frecuencias mayores que 200 ocurrencias por t_{ancho} ($f > 0.2$Hz), mientras que en el caso de la muerte cerebral, el espectro consiste de unos picos delgados. Series de tiempo fractales tienen característicamente espectros de banda ancha [49].

En la cuarta fila, se presentan los espectros de potencia en escala logarítmica. Los picos individuales en el espectro en escala lineal muestran un comportamiento lineal, es decir, cómo una enfermedad o un tratamiento influye en la potencia P o la posición f de un pico individual a una determinada frecuencia. En cambio, en escala logarítmica, se disimulan los picos individuales del espectro de potencias y se pone en evidencia el comportamiento global del conjunto de todas las frecuencias. En los tres casos, a bajas frecuencias (ritmos menores de 100 ocurrencias por t_{ancho} o ritmos menores de una vez por 10s), se observa que el espectro de potencias se comporta como una ley de potencias $P(f) \sim f^{\beta}$ (nótese que el pico de la respiración no hace parte del comportamiento de ley de potencias [31, 32]). Encontramos $\beta \approx -0.35$ en el caso de la enfermedad coronaria (entre el ruido blanco y el ruido $1/f$), $\beta \approx -0.92$ en el caso sano (muy cerca del ruido $1/f$), y $\beta \approx -2.2$ para el caso de la muerte cerebral (muy cerca del ruido browniano). Las series cardíacas en las cuales se basa este análisis son relativamente cortas, de unos 15 minutos (véase la Tabla 2), así que se puede seguir el comportamiento de ley de potencias solamente hasta ritmos de $f = (15\text{min})^{-1}$. En estudios de 24 horas con un holter, se puede seguir tales leyes de potencia hasta ritmos circadianos. Otra ventaja de series de tiempo más largas es que se podría partir la serie larga en fragmentos más cortos, de los cuales se podrían calcular los espectros de potencias, y luego el espectro de potencias promedio sobre todos los fragmentos. El resultado es que se muestra el comportamiento de ley de potencias más claramente, con menos fluctuaciones [37]. Se muestra también la frecuencia promedia $\langle f \rangle$ (véase la ec. (5)) de la parte del espectro de potencias que se comporta como ley de potencias, que en la sección siguiente se utilizará para calcular el tiempo propio de la serie de tiempo.

En la última fila, se muestran las funciones de autocorrelación $C(\tau)$ de las tres series de tiempo. La función de autocorrelación se puede calcular con su fórmula matemática a partir de la serie de tiempo, véase la ec. (4), o utilizando el teorema de Wiener-Khinchin con la transformada de Fourier inversa (FFT^{-1}) del espectro de potencias [37, 43]. Aquí se calculó la función de autocorrelación con FFT^{-1} del espectro de potencias completa, $C(\tau)_{\text{compl}}$ (gris), y de la parte del espectro de potencias que se comporta como una ley de potencias, $C(\tau)_{\text{parte}}$ (negro). Los dos cálculos de la función de autocorrelación corresponden, $C(\tau)_{\text{compl}} \approx C(\tau)_{\text{parte}}$, con la excepción que $C(\tau)_{\text{parte}}$ fluctúa más lentamente porque no incluye la parte de altas frecuencias del espectro de potencias. Cuantificamos la complejidad de la serie de tiempo con el alcance de sus correlaciones, lo que llamamos la memoria. La memoria es el tiempo τ_0 después del cual la función de autocorrelación decae hasta 0, $C(\tau_0)_{\text{parte}} = 0$. Encontramos que en el sujeto sano la complejidad y la memoria se maximizan, $\tau_0 = 60$, mientras que en el caso de la enfermedad coronaria y la muerte cerebral la complejidad es menor, $\tau_0 = 8$ y $\tau_0 = 48$, respectivamente. En la sección siguiente, encontraremos que es mejor expresar la memoria en el tiempo propio de la serie de tiempo, con el factor de renormalización

$\langle t \rangle$, véase la ec. (6). En estas unidades se obtiene una complejidad máxima para el sujeto sano, $\tau_0/\langle t \rangle = 2.03$, y $\tau_0/\langle t \rangle = 1.53$ y $\tau_0/\langle t \rangle = 0.71$ para la enfermedad coronaria y la muerte cerebral, respectivamente.

Conclusión. Las serie del sujeto sano es más compleja que las series que corresponden con la muerte cerebral y la enfermedad coronaria, las cuales aparentemente perdieron complejidad según la hipótesis de Lipsitz y Goldberger. Resulta también que se puede perder complejidad de dos maneras, (i) cuando la serie se degenera hacia el ruido blanco aleatorio ($\beta = 0$), o (ii) cuando la serie se vuelve demasiado regular y rígida como en el caso del ruido browniano ($\beta = -2$). Parece que se confirma el diagrama de fase de la complejidad en función de la fuerza de correlación β (Figura 3). En un protocolo de investigación experimental se podría proponer comparar el diagrama de fase de complejidad con el diagrama de adaptabilidad de Chrousos (Figura 1), para ver si todas las enfermedades relacionadas con respuestas deficientes/excesivas al estrés corresponden con series aleatorias/rígidas, y vice versa.

TABLE 2. Cambio de unidades de la frecuencia f del espectro de potencias de la Fig. 5 de ocurrencias por duración de la serie de tiempo a las unidades Hz y latidos por minuto (bpm). t_{pix} y t_{ancho} dan respectivamente la resolución y la duración de la serie de tiempo, y f_{pix} y f_{ancho} la resolución y el ancho de la primera mitad del espectro de potencias.

	Enfermedad coronaria	Sano	Muerte cerebral
$t_{\text{pix}} = \langle \Delta t \rangle$ $t_{\text{ancho}} = 1000 \langle \Delta t \rangle$	842ms 842s=14min	1067ms 1067s=17.8min	606ms 606s=10min
$f_{\text{pix}} = 1/t_{\text{ancho}}$ $f_{\text{ancho}} = 0.5/t_{\text{pix}}$	1.2mHz=0.071bpm 0.59Hz=35.64bpm	0.94mHz=0.056bpm 0.47Hz=28.11bpm	1.7mHz=0.099bpm 0.83Hz=49.54bpm

Simulación teórica

El generador de ruidos

El propósito de esta sección es averiguar el diagrama de fase de complejidad en función de la fuerza de correlación β (Fig. 3(b)) con una simulación numérica en la cual generamos ruidos fractales que obedecen la ley de potencias $P(f) \sim f^{\beta}$ de la ec. (1) con exponente espectral $-10 \leq \beta \leq 10$, para luego calcular la memoria y la complejidad correspondientes. Utilizamos el generador de ruidos "filtro FFT" [37, 43, 50],

$$F_{\beta}(t_0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \; e^{i\omega \, t_0} \; \omega^{\beta/2} \left(\int_{-\infty}^{+\infty} dt \; F_0(t) \; e^{-i\omega t} \right), \tag{3}$$

lo cual genera una serie de tiempo fractal $F_{\beta}(t)$ con el exponente espectral β deseado. El generador toma la transformada de Fourier inversa del ruido blanco aleatorio $F_0(t)$, y lo deja pasar por el filtro $\omega^{\beta/2}$ adecuado. En esta simulación, las series generadas tienen una dimensión dim $= 20001$, fluctúan alrededor de 0 en el rango $[-1, 1]$, y se pueden apreciar en las Refs. [37, 43]. Suponemos que las series representan secuencias de intervalos de tiempo $\Delta t(i)$ entre latidos sucesivos en el corazón. En la Fig. 6, se

presentan los espectros de potencia correspondientes. *Series correlacionadas* tienen un efecto de amontonamiento (clustering en inglés), lo que significa que cuando un valor de la serie de tiempo es mayor (menor) que el promedio, la probabilidad es muy alta que los valores vecinos también serán mayores (menores) que el promedio [51]. En cambio, *series anti-correlacionadas* exhiben un efecto de anti-amontonamiento, donde la probabilidad es muy alta que valores mayores y menores que el promedio se alternan [51]. La probabilidad del efecto de (anti-) amontonamiento es mayor para fuerzas de correlación más grandes $|\beta|$ pero tiende a perderse entre valores más distantes de la serie de tiempo, véase la Fig. 7.

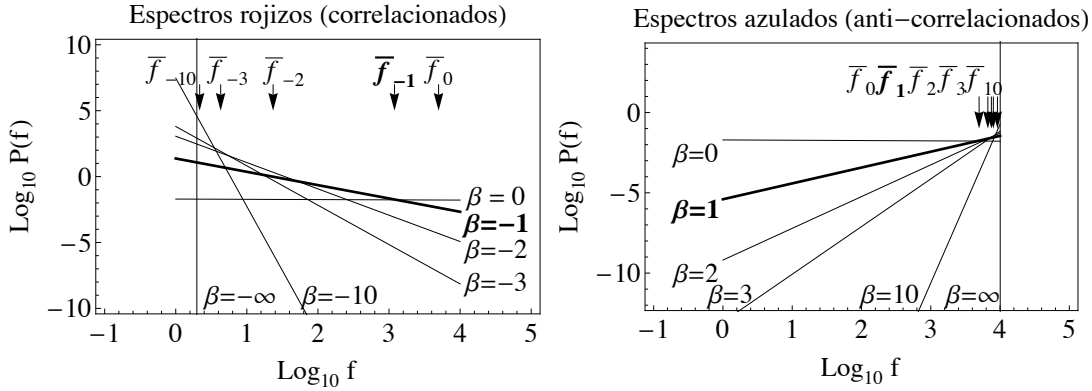


FIGURE 6. Espectros de potencia rojizos ($\beta \leq 0$) y azulados ($\beta \geq 0$) corresponden con series de tiempo correlacionadas y anticorrelacionadas, respectivamente. Están indicadas también las frecuencias promedias $\langle f \rangle$ de los diferentes espectros de potencia, véase la ec. (5).

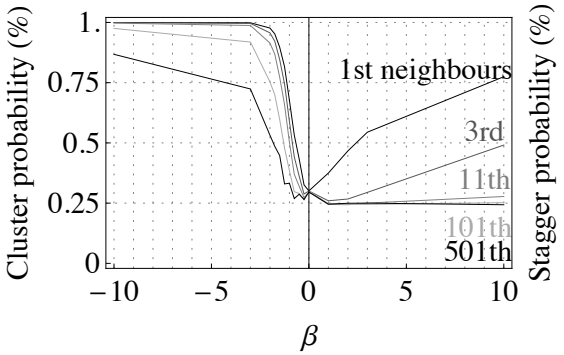


FIGURE 7. Probabilidad de (anti-) amontonamiento en series de tiempo (anti-) correlacionadas. Para series correlacionadas ($\beta \leq 0$), se muestra la probabilidad de que si un valor de la serie de tiempo es mayor (menor) que el promedio, su primer vecino también es mayor (menor) que el promedio. Se muestra tal probabilidad también para los vecinos a distancias 3, 11, 101 y 501. Para series anti-correlacionadas ($\beta \geq 0$), se muestra la probabilidad de que si un valor de la serie de tiempo es mayor (menor) que el promedio, su primer vecino es menor (mayor) que el promedio. Se muestra tal probabilidad también para los vecinos a distancias 3, 11, 101 y 501. Series de tiempo más correlacionadas que el ruido browniano tienen un comportamiento de amontonamiento que es casi perfecto y que perdura muchos cientos de valores. Entre el ruido browniano y el ruido blanco, la eficiencia del efecto de amontonamiento se comporta lineal con la fuerza de correlación $|\beta|$. La eficiencia del efecto de alternancia en series anti-correlacionadas también se comporta lineal con la fuerza de anti-correlación $|\beta|$, pero el efecto de alternancia se pierde rápidamente con la distancia entre vecinos.

La función de autocorrelación y la memoria

El paso siguiente es calcular la *función de autocorrelación*,

$$\begin{aligned} C(\tau) &= \frac{\sum_{i=1}^{N-\tau}\left(\Delta t(i)-\langle\Delta t\rangle\right)\left(\Delta t(i+\tau)-\langle\Delta t\rangle\right)}{\sum_{i=1}^{N-\tau}\left(\Delta t(i)-\langle\Delta t\rangle\right)^2} \\ &= \frac{\sum_{i=1}^{N-\tau}\left(\Delta t(i)-\mu\right)\left(\Delta t(i+\tau)-\mu\right)}{\sigma^2}, \end{aligned} \tag{4}$$

donde μ y σ son el promedio y la desviación estándar de la serie de tiempo, respectivamente. Representamos en la Fig. 8 las funciones $C(\tau)$ obtenidas para las series con diferentes β. En el panel (a) observamos que mientras más correlacionada la serie de tiempo, más lentamente que $C(\tau)$ decae. La función de autocorrelación del ruido $1/f$ parece no tener una memoria excepcional. Sin embargo, mientras más correlacionada la serie de tiempo, más que las frecuencias bajas dominan el espectro. Uno puede calcular la *frecuencia promedia* del espectro,

$$\langle f\rangle = \frac{\sum_{f=f_{\min}}^{f_{\max}} f\,P(f)}{\sum_{f=f_{\min}}^{f_{\max}} P(f)}. \tag{5}$$

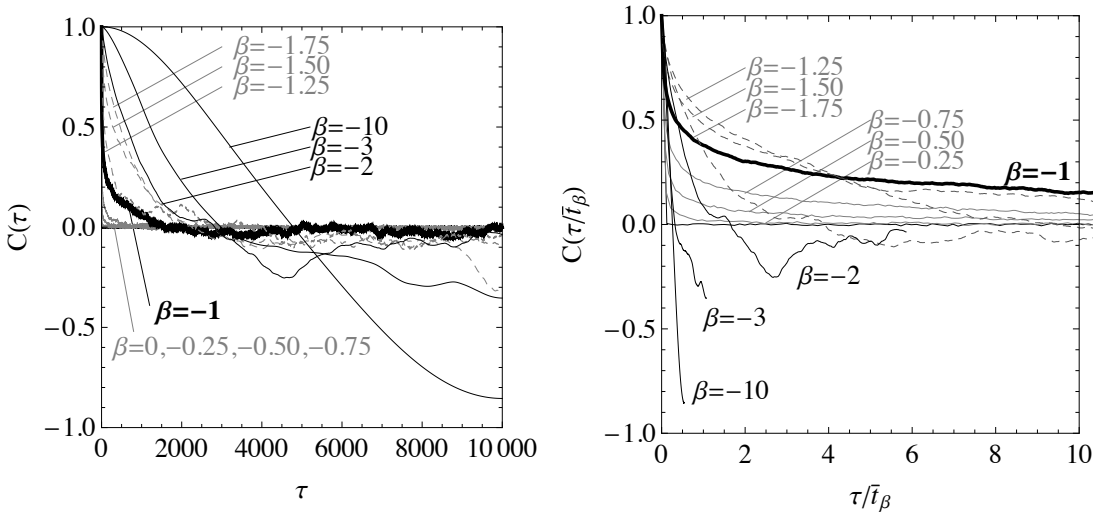
Se puede observar que esta frecuencia promedia se mueve hacia las frecuencias más bajas para las series más correlacionadas (véanse las flechas en la Fig. 6). A parte de la frecuencia promedia del espectro de potencias, también la frecuencia promedia del espectro de amplitudes podría servir el mismo propósito [52]. En consecuencia de lo anterior, mientras más correlacionada la serie de tiempo, más que son los ritmos lentos que la dominan. Se puede calcular el *período promedio* de la serie de tiempo,

$$\langle t\rangle = \frac{\dim}{2\,\langle f\rangle}. \tag{6}$$

En el *tiempo absoluto* τ de la Fig. 8(a), las series más correlacionadas pueden bien gozar de una memoria más larga, sin embargo, esta memoria debe de ser interpretada en relación con los ritmos típicos de la serie de tiempo, es decir, en relación con el *tiempo propio* de la serie de tiempo, $\tau/\langle t\rangle$. De esta manera, en la Fig. 8(b), el ruido $1/f$ surge como la serie de tiempo que se decae más lentamente. También en estudios con circuitos electrónicos se obtiene que la función de autocorrelación decae más lentamente para el ruido $1/f$ [53]. En la Fig. 9, se muestran la frecuencia promedia $\langle f\rangle$ y el período promedio $\langle t\rangle$, los cuales se utilizan para renormalizar el tiempo, en función de la fuerza de correlación $|\beta|$. En la Fig. 10, se muestra la memoria en función de la fuerza de correlación, en tiempo absoluto τ_0 (panel (a)) y en el tiempo propio de la serie de tiempo $\tau_0/\langle t\rangle$ (panel (b)).

Conclusión. En esta simulación numérica obtenemos que la complejidad de la serie de tiempo - cuantificada como su memoria - se maximiza para el ruido $1/f$ cuando se estudia en el tiempo propio de la serie de tiempo. Como en el esquema de la Fig. 3(b),

se pierde complejidad cuando la serie de tiempo se degenera desde el ruido $1/f$ hacia el ruido blanco o hacia el ruido browniano. En esta simulación, se puede observar que pequeñas desviaciones del punto crítico $\beta = -1$ inducen drásticas pérdidas de complejidad. Es este un efecto no-lineal, que implica algo en la fisiología?



(a) Tiempo absoluto τ

(b) Tiempo propio $\tau/\langle t \rangle$

FIGURE 8. La función de autocorrelación (a) en tiempo absoluto τ, y (b) en el tiempo propio de cada serie de tiempo $\tau/\langle t \rangle$. En el tiempo absoluto τ, parecería que mientras mayor la fuerza de correlación de la serie de tiempo $|\beta|$, mayor el tiempo que necesita la función de autocorrelación para decaer hasta 0 (la memoria). Sin embargo, en el tiempo propio de cada serie de tiempo, $\tau/\langle t \rangle$, se puede apreciar que es el ruido $1/f$ que decae más lentemente y que es el ruido $1/f$ donde el efecto de memoria se maximiza.

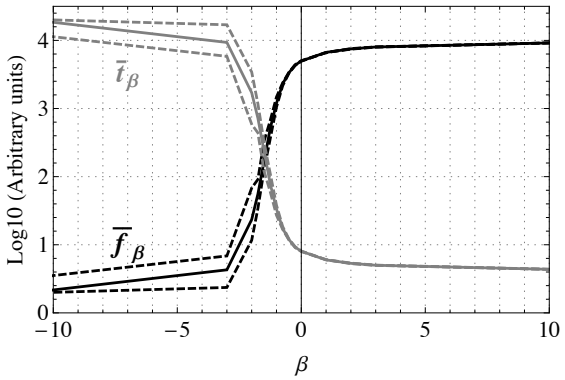
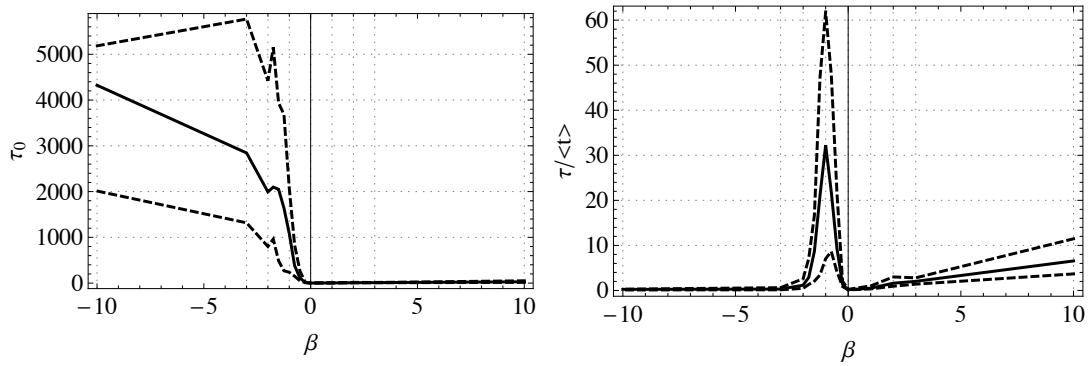


FIGURE 9. Frecuencia promedia $\langle f \rangle$ (negro) y período promedio $\langle t \rangle$ (gris) en función de la fuerza de correlación $|\beta|$. Para cada β se generó un ensamble de 10 realizaciones de la serie de tiempo. Para cada β, se muestra el mínimo (linea quebrada), promedio (linea en negrita), y máximo (linea quebrada) de $\langle f \rangle$ y $\langle t \rangle$ sobre el ensamble.



(a) Tiempo absoluto τ_0 (b) Tiempo propio $\tau_0/\langle t\rangle$

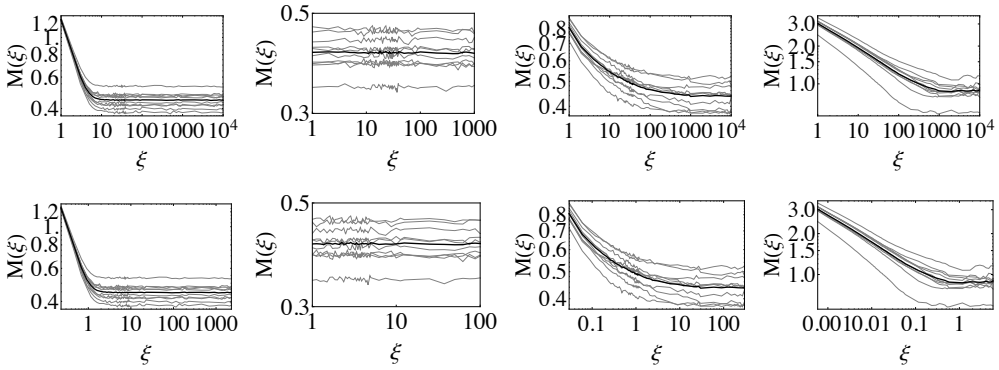
FIGURE 10. Evolution of the memory τ_{coh} with the (anti)correlation strength β. The memory is taken as the time τ it takes for the autocorrelation function to drop below a certain threshold value $|C(\tau)| < 0.50$ or 0.10. For high threshold values around 0.5 (HWHM), memory is maximized for $\beta = -1.50$, but for lower thresholds the memory maximizes for $1/f$ noise and drops quickly both for rising or diminishing correlation strength $|\beta|$. The memory drops to zero quickly for $\beta \to 0$, but starts to rise slowly and linearly for rising anti-correlation strength. The memory falls to almost zero for $\beta - > -2$ and continues to drop to zero for rising correlation strength.

La función de información mutua

Muchas veces, series no-estacionarias constituyen un problema para un análisis estadístico. Una serie es no-estacionaria cuando sus momentos (promedio, varianza, etc.) no son definidos [54]. En consecuencia, para series no estacionarias no se puede definir la función de autocorrelación $C(\tau)$ de la ec. 4. No siempre es fácil estar seguro si una serie es estacionaria o no. Del otro lado, se puede calcular la *función de información mutua*, $M(\xi)$, que tiene un comportamiento similar. La función de información mutua es aplicable a series no-estacionarias, y se puede utilizar para estudiar correlaciones generales entre dos variables [55], donde la función de autocorrelacón sólo muestra correlaciones lineales. En la Fig. 11, se muestran las funciones $M(\xi)$ para ensambles de 10 realizaciones de series para $\beta = 10, 0, -1, -2$ en tiempo absoluto y tiempo propio. En la Fig. 12, se muestra la memoria en función de la fuerza de correlación. Los resultados son similares a los obtenidos con la función de autocorrelación.

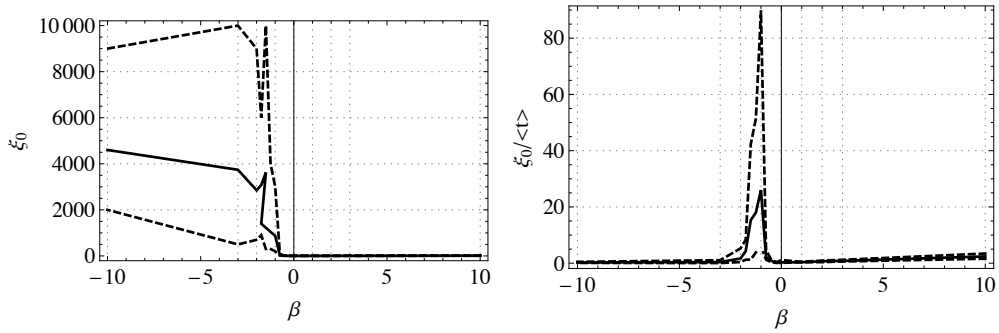
Otras medidas

En la Fig. 13 se presenta el comportamiento de la medida de la *Entropía Aproximada* ApEn en función de la correlación. La ApEn cuantifica la impredecibilidad de la serie de tiempo [56], y en la figura se puede apreciar que la ApEn maximiza para el ruido blanco. Cuando Vaillancourt y Newell se basaron en la ApEn para cuantificar la complejidad y llegaron a la conclusión que la complejidad puede aumentar en condiciones adversas de salud, no estaban midiendo un aumento de complejidad pero un aumento en lo aleatorio de la serie de tiempo [44, 45, 46].



(a) $\beta = +10$ (b) $\beta = 0$ (c) $\beta = -1$ (d) $\beta = -2$

FIGURE 11. Función de información mutua $M(\xi)$ en escala log-log para series de tiempo con (a) $\beta = 10$, (b) $\beta = 0$, (c) $\beta = -1$ y (d) $\beta = -2$. Para cada β, se muestra $M(\xi)$ para 10 realizaciones diferentes de la serie de tiempo (curvas grises) y el comportamiento promedio $\langle M(\xi)\rangle$ sobre el ensamble (curvas en negrita). Se muestra $M(\xi)$ en tiempo absoluto ξ (paneles superiores) y en el tiempo propio $\xi/\langle t\rangle$ para cada β (paneles inferiores).



(a) Memoria ξ (tiempo absoluto) (b) Memoria $\xi/\langle t\rangle$ (tiempo propio)

FIGURE 12.

El *exponente de Hurst H* mide la velocidad de difusión. En la Fig. 14, se muestra la difusión de un ensamble de 10 realizaciones de ruido $1/f$ en comparación con la difusión de un ensamble de 10 series brownianas. La difusión de una "nuve" de series brownianas sigue la ley $\sigma \propto t^{1/2}$, la cual se puede generalizar a $\sigma \propto t^H$, donde cada serie fractal tiene su exponente de Hurst característico [55, 57]. En la fig. 15, se muestra el comportamiento del exponente de Hurst en función de la fuerza de correlación β.

La dimensión fractal D de una serie de tiempo se puede medir con el método de conteo de cajas [58] y con un mejorado método de conteo de cajas [59, 60]. Una sola caja Δt que se desliza sobre la serie de tiempo resulta en un problema unidimensional con una dimensión fractal $D_{1D} \in [0,1]$. Se define una longitud en función de la resolución en el tiempo Δt,

$$L_{1D} = \frac{1}{\Delta t}\sum_{i=1}^{N} |F(i+\Delta t) - F(i)|\,. \tag{7}$$

Del otro lado, dividir la serie de tiempo en cajas de igual tamaño resulta en un problema bidimensional con una dimensión fractal $D_{2D} \in [1,2]$. Encontramos para la longitud,

$$L_{2D} = \frac{1}{\Delta t} \sum_{i=2}^{N/\Delta t} \left| F\left[i\,\Delta t\right] - F\left[(i-1)\,\Delta t\right] \right|. \tag{8}$$

La dimensión fractal se define con

$$D = -\frac{d \log L(\Delta t)}{d log \Delta t}. \tag{9}$$

En la fig. 15, se muestra el comportamiento de las dimensiones fractales D_{1D} y D_{1D} en función de la fuerza de correlación β.

El Detrended Fluctuation Analysis (DFA) se parece mucho al análisis de Hurst, sólo que aquí antes de hacer el análisis se quitan las tendencias de la serie de tiempo [61]. En la fig. 16, se muestra el comportamiento de las dimensiones fractales D_{1D} y D_{1D} en función de la fuerza de correlación β.

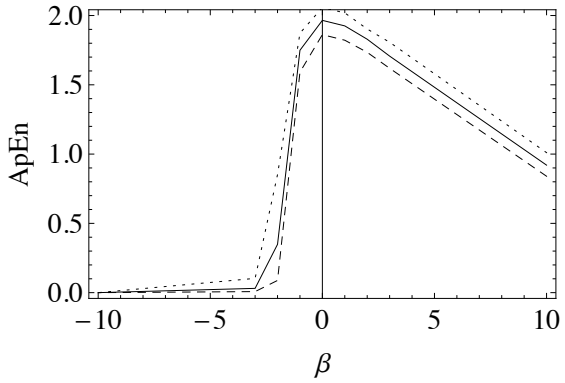


FIGURE 13. La entropía aproximada ApEn(m,r,N) con $m = 2, r = 0.20\sigma$, y $N = 1000$ en función de la fuerza de correlación β. La linea continua es el promedio de la entropía de un ensamble de 200 subseries de dimensión N, y las lineas quebradas corresponden con el mínimo y el máximo de la entropía del ensamble. ApEn mide la regularidad, donde ApEn=0 para series periódicas y ApEn≈ 2 para series aleatorias.

ACKNOWLEDGMENTS

Se agradece al Dr. J.C. López Vieyra por ayuda con la programación del generador de ruidos y discusiones; al Dr. A. Frank por discusiones; al Dr. de Godoy por los datos cardíacos; a los organizadores del congreso "III Jornada de Aplicações Médicas de Fractais, Caos e Complexidade" (15 de octubre de 2011, Hospital Estadal de Bauru, São Paulo, Brasil) y sobretodo al Dr. A. Sant'Ana por la posibilidad de presentar este trabajo de investigación; al Dr. R. Mansilla por discusiones sobre la (no-)estacionariedad de series de tiempo.

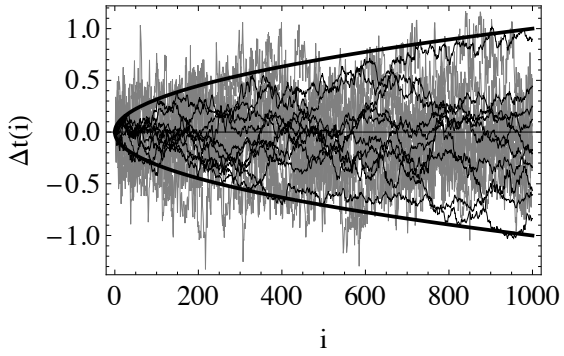


FIGURE 14. El escalamiento de una nuve de partículas brownianas en difusión. Se muestran las trazas de un ensamble de 10 realizaciones de series brownianas (negro), en comparación con 10 realizaciones del ruido $1/f$. Todas las realizaciones empiezan en el mismo punto para tiempo $t = 0$. Se puede apreciar que para los ruidos $1/f$ casi no hay dispersión, mientras que las series brownianas se difunden según la ley $\sigma \propto t^{1/2}$ (curva en negrita). Figura adaptada de [55].

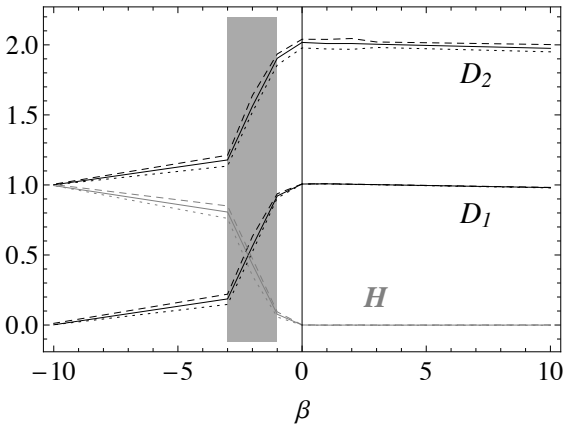


FIGURE 15. La dimensión fractal D y el exponente de Hurst H en función de la fuerza de correlación β. La dimensión fractal se calculó con el método de conteo de cajas (box counting), donde hay dos posibilidades según el planteamiento de un problema unidimensional (ec. 7) o bidimensional (ec. 8). Los resultados se diferencian en una sola unidad. Las relaciones $D = 2-H$, $H = -(\beta+1)/2$ y $D = (\beta+5)/2$ son válidos sólo en el rango $-3 < \beta < -1$ (área gris). Se puede apreciar que fuera de este rango, D y H se saturan rápidamente, de un lado $D, H \to 1$ para $\beta < -3$, y del otro lado $D \to 2$ y $H \to 0$ para $\beta > -1$. Nótese que la dimensión fractal o el exponente de Hurst (casi) no puede distinguir entre el ruido blanco y el ruido $1/f$.

REFERENCES

1. A. Omran, *Milbank Mem Fund Q* **49**, 509–538 (1971).
2. L. M. Gutiérrez Robledo, *J. Gerontol.* **57A**, M162–M167 (2002).
3. A. Beswick, K. Rees, P. Dieppe, S. Alvis, et al., *Lancet* **371**, 725–735 (2008).
4. M. C. Mackey, and L. Glass, *Science* **197**, 287–289 (1977).
5. M. C. Mackey, and J. G. Milton, *Ann. NY Acad. Sci.* **504**, 16–32 (1987).
6. L. Glass, A. Beuter, and D. Larocque, *Math. Biosci.* **90**, 111–125 (1988).
7. L. Glass, *Nature* **410**, 277–284 (2001).
8. A. C. Ahn, M. Tewari, C.-S. Poon, and R. S. Phillips, *PLoS Medicine* **3**, e208 (2006), 0709–0713.

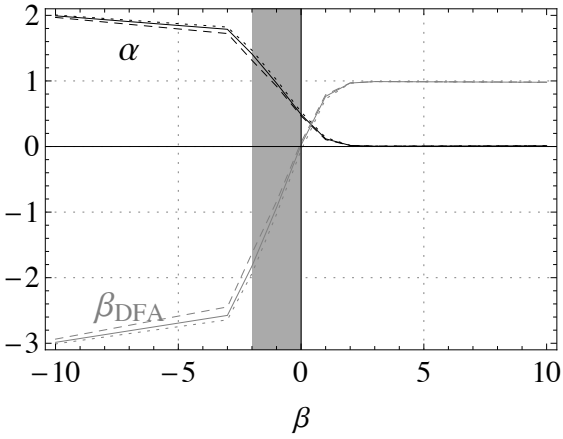


FIGURE 16. El valor del exponente α del DFA en función de la fuerza de correlación β. El valor de la fuerza de correlación se puede estimar con el DFA como $\beta_{DFA} = 2\alpha - 1$. Se puede apreciar que la fuerza de correlación estimada con el DFA se satura cerca de $\beta_{DFA} = -3$ y la fuerza de anti-correlación cerca de $\beta_{DFA} = 1$. El método DFA debería aplicarse sólo al rango $-2 \leq \beta \leq$, fuera del cual el valor del exponente α_{DFA} se satura rápidamente y no logra distinguir entre diferentes fuerzas de (anti-) correlación.

9. A. C. Ahn, M. Tewari, C.-S. Poon, and R. S. Phillips, *PLoS Medicine* **3**, e209 (2006), 0956–0960.
10. W. Herfel, D. Rodrigues, and Y. Gao, *J. Chin. Philosophy* **34 suppl.1**, 57–79 (2007).
11. S. Chang, *Chin. J. Physiol.* **53** (2010).
12. D. Lloyd, M. A. Aon, and S. Cortassa, *The Scientific World* **1**, 133–145 (2001).
13. F. E. Yates, *Ecol. Psychol.* **20**, 148–179 (2008).
14. A. L. Goldberger, D. R. Rigney, and B. J. West, *Sci. Am.* **262**, 34–41 (1990).
15. T. G. Buchman, *Nature* **420**, 246–251 (2002).
16. G. P. Chrousos, and P. W. Gold, *J. Am. Med. Assoc.* **267**, 1244–1252 (1992).
17. G. P. Chrousos, *Nat. Rev. Endocrinol.* **5**, 374–381 (2009).
18. T. G. Guilliams, and L. Edwards, *The Standard* **9** (2010).
19. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, *Eur. Heart J.* **17**, 354–381 (1996), published simultaneously in *Circulation* 93 (1996) 1043-1065.
20. E. Landa, R. Fossion, I. O. Morales, C. Hernández, V. Veláquez, J. C. López Vieyra, and A. Frank, *Rev. Mex. Fís* **S55**, 50–59 (2009).
21. E. Landa, I. O. Morales, R. Fossion, P. Stránský, V. Veláquez, J. C. López Vieyra, and A. Frank, *Phys. Rev. E* **84**, 016224 (5 pages) (2011).
22. M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker, *Nature* **413**, 591–596 (2001).
23. M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara, *Nature* **461**, 53–59 (2009).
24. S. Ahmad, T. Ramsay, L. Huebsch, S. Flanagan, S. McDiarmid, I. Batkin, L. McIntyre, S. R. Sundaresan, D. E. Maziak, F. M. Shamji, P. Hebert, D. Fergusson, A. Tinmouth, and A. J. E. Seely, *PLoS One* **4**, e6642 (2009), (10 pages).
25. *Manual de uso*, VECCSA S.A., Argentina (????), URL `www.cardiovex.com.ar`, cardioVex, Registradores Holter MMC10/MMC10D.
26. L. Lipsitz, *J. Neurobiol.* **57A**, B115–B125 (2002).
27. J. C. Nelson, Rizwan-Uddin, M. P. Griffin, and J. R. Moorman, *Pediatric Res.* **43**, 823–831 (1998).
28. R. Pool, *Science* **243**, 604–607 (1989).
29. L. Glass, *J. Cardiovasc. Electrophysiol.* **10**, 1358–1360 (1999).
30. L. Glass, *Chaos* **19**, 028501 (2009).
31. M. Kobayashi, and T. Musha, *IEEE Trans. Biomed. Engin.* **BME-29**, 456–457 (1982).
32. D. T. Kaplan, and M. Talajic, *Chaos* **1**, 251–256 (1991).

33. I. Prigogine, *As leis do caos*, Fundação Editora Da UNESP, São Paulo, Brasil, 2000 (Titulo original em italiano: "Le leggi del caos", 1993).
34. M. Gell-Mann, *Complexity* **1**, 16–19 (1995).
35. P. Bak, *How Nature works: The science of self-organized criticality*, Springer-Verlag, New York, 1996.
36. T. Gisiger, *Biol. Rev.* **76**, 161–209 (2001).
37. R. Fossion, "Fractales en la Medicina y en la Geriatría: Patrones de orden en series de tiempo fisiológicas," in *Memorias de la XVIII Escuela de Verano en Física, 26 julio – 6 agosto, 2010*, edited by R. Jáuregui, J. Récamier, M. Torres, and R. Pérez Pascual, Instituto de Física (UNAM), Instituto de Ciencias Físicas (UNAM), México D.F., Cuernavaca, 2011, pp. 89–110.
38. J. Hayano, Y. Sakakibara, M. Yamada, N. Ohte, T. Fujinami, K. Yokoyama, Y. Watanabe, and K. Takata, *Circulation* **81**, 1217–1224 (1990).
39. J. Hayano, F. Yamasaki, S. Sakata, A. Okada, S. Mukai, and T. Fujinami, *Am. J. Physiol. Heart Circ. Physiol.* **273**, 2811–2816 (1997).
40. S. M. Pikkujämsä, T. H. Mäkikallio, L. B. Sourander, I. J. Räihä, P. Puukka, J. Skyttä, C.-K. Peng, A. L. Goldberger, and H. V. Huikuri, *Circulation* **100**, 393–399 (1999).
41. L. A. Lipsitz, and A. L. Goldberger, *J. Am. Med. Assoc.* **267** (1992).
42. R. Fossion, "Una definición 'compleja' de la fragilidad: Caos, fractales y complejidad en series de tiempo biológicas," in *Envejecimiento humano: Una visión transdisciplinaria*, edited by L. M. Gutiérrez Robledo, and J. H. Gutiérrez Ávila, Instituto de Geriatría, México D.F., 2010, chap. XVII, pp. 171–183, ISBN 978-607-460-121-3.
43. R. Fossion, "Scale invariance as a symmetry in physical and biological systems: Listening to photons, bubbles and heartbeats," in *Symmetries in Nature: Symposium in Memoriam Marcos Moshinsky (Cuernavaca, Mexico, 7 – 14 august*, edited by L. Benet, P. O. Hess, J. M. Torres, and K. B. Wolf, AIP Conf. Proc., New York, 2010, vol. 1323, pp. 74–90, ISBN 978-0-7354-0877-7.
44. D. E. Vaillancourt, and K. M. Newell, *Neurobiol. Aging* **23**, 1–11 (2002).
45. D. E. Vaillancourt, and K. M. Newell, *Neurobiol. Aging* **23**, 27–29 (2002).
46. A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, *Neurobiol. Aging* **23**, 23–26 (2002).
47. Polar mexico (????), URL `http://www.polar.com.mx/mx-es`, monitores de frecuencia cardíaca desde 1977.
48. M. F. de Godoy, Teoria do caos aplicada à medicina, Faculdade de Medicina da Universidade de São José do Rio Preto (FAMERP), São Paulo, Brasil (2003), URL `http://www.mfgodoy.med.br/caos.pdf`, tese apresentada para obtenção do título de Livre Docente em Cardiologia.
49. A. L. Goldberger, D. R. Rigney, J. Mietus, E. Antman, and S. Greenwald, *Experientia* **44**, 983–987 (1988).
50. W. H. Press, *Comments Astrophys.* **7**, 103–119 (1978).
51. C.-K. Peng, J. Mietus, J. Hausdorff, S. Havlin, H. Stanley, and A. Goldberger, *Phys. Rev. Lett.* **70**, 1343–1346 (1990).
52. M. T. Rosenstein, J. J. Collins, and C. J. D. Luca, *Phys. D* **65**, 117–134 (2001).
53. M. S. Keshner, *Proc. IEEE* **70**, 212–218 (1982).
54. L. S. Liebovitch, and D. Scheurle, *Complexity* **5**, 34–43 (2000).
55. P. S. Addison, *Fractals and chaos: An illustrated course*, Institute of Physics Publishing, Bristol, Philadelphia, ????
56. S. M. Pincus, and A. L. Goldberger, *Am. J. Physiol.* **266**, H1643–H1656 (1994).
57. Z.-Y. Su, and T. Wu, *Phys. A* **380**, 418–428 (2007).
58. I. McHardy, and B. Czerny, *Nature* **325**, 696–698 (1987).
59. T. Higuchi, *Physica D* **31**, 277–283 (1988).
60. T. Higuchi, *Physica D* **46**, 254–264 (1990).
61. A. L. Goldberger, L. A. N. Amaral, J. M. Hausdorff, P. C. Ivanov, C.-K. Peng, and H. E. Stanley, *Proc. Nat. Acad. Sci.* **99**, 2466–2472 (2002).

Cosmología Moderna

G. Germán*

Instituto de Ciencias Físicas,

Universidad Nacional Autónoma de México,

Apdo. Postal 48-3, 62251 Cuernavaca, Morelos, México

1. Introducción

La cosmología moderna está apoyada principalmente en cuatro tipos de observaciones: la expansión (de Hubble) del universo, la radiación cósmica de microondas de fondo (con una temperatura de $2{,}725K$), la nucleosíntesis de elementos ligeros que cuantifica la abundancia cósmica de hidrógeno, helio, deuterio, trazas de litio y berilio, y la formación de galaxias y estructura a grandes escalas.

Típicamente, las observaciones cosmológicas se realizan usando el espectro electromagnético completo. En orden creciente de energías se dispone de ondas de radio, microondas, infrarrojo, luz visible, ultravioleta, rayos-x y rayos gama. Sin embargo más recientemente se han realizado observaciones utilizando neutrinos (que son partículas que interactúan débilmente) y rayos cósmicos (partículas altamente relativistas). Las observaciones a muy grandes escalas (a nivel de supercúmulos de galaxias) apoyan el Principio Cosmológico, que es la proposición de que el universo es homogéneo e isotrópico. Este principio encuentra su expresión matemática a través de la métrica de Friedmann-Robertson-Walker (FRW). Otro aspecto de vital importancia al que han llevado las observaciones es el de la expansión del universo. Prácticamente todas las galaxias se están alejando de nosotros lo cual se expresa mediante el corrimiento al rojo z definido por $z = \delta\lambda/\lambda_0 = \lambda_{obs}/\lambda_0 - 1$.

Hubble descubrió también que cuanto mas alejada se encuentra una galaxia mayor será su velocidad de recesión. Esto queda establecido mediante la ley de Hubble $v = H_0 r$, la cual nos lleva al concepto de la gran explosión (Big-Bang). La cosmología observacional también permite estudiar la composición del universo. El contenido de materia del universo observado o conjeturado está dado por bariones, radiación, neutrinos, materia obscura y energía obscura.

Por la parte teórica la cosmología descansa básicamente en el Principio Cosmológico, las ecuaciones de la Relatividad General de Einstein y la ecuación de estado del tipo de materia

*Correo-e: gabriel@fis.unam.mx.

bajo estudio. El Principio Cosmológico es la hipótesis de que todos los puntos del universo son equivalentes. Este principio se implementa a través de la métrica FRW para un universo homogéneo e isótropo. En este caso, el intervalo espacio-temporal está dado por

$$ds^2 = a(t)^2 \left(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \right), \tag{1}$$

en donde $a(t)$ es el factor de escala, o radio, del universo y es una función del tiempo. Su evolución nos da una medida de la expansión de éste. En la Eq.(1) arriba, k es una constante relacionada con la curvatura del espacio-tiempo. Las ecuaciones de Einstein son ecuaciones entre tensores que relacionan la distribución de energía-momento de la materia con la geometría del espacio-tiempo

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu} + \Lambda g_{\mu\nu}. \tag{2}$$

De la Eq.(1) podemos leer directamente la métrica de FRW

$$g_{\mu\nu} = diag\left(-1, \frac{a^2(t)}{1-kr^2}, a^2(t) r^2, a^2(t) r^2 \sin^2\theta \right), \tag{3}$$

en tanto que el tensor de energía-momento para un fluído perfecto es

$$T_{\mu\nu} = diag\left(\rho, -p, -p, -p\right). \tag{4}$$

La ecuación de estado establece una relación entre la presión y la densidad de energía $p = p(\rho)$ y está dada por

$$p = \omega\rho. \tag{5}$$

De las ecuaciones de Einstein para $g_{\mu\nu}$ y $T_{\mu\nu}$ dadas arriba se sigue la ecuación de Friedmann

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{k}{a^2}, \tag{6}$$

y la ecuación de aceleración

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + 3p\right). \tag{7}$$

La ecuación de continuidad o de conservación de la energía se deriva de las últimas dos

$$\dot{\rho} = -3H\left(\rho + p\right), \tag{8}$$

donde H es la "constante" de Hubble definida por

$$H = \frac{\dot{a}}{a}. \tag{9}$$

La constante k que aparece en la Eq.(1) determina la curvatura del universo. Si $k = 0$ el universo es plano, en tanto que si $k > 0$ o $k < 0$ el universo es esférico (cerrado) o hiperbólico (abierto), respectivamente. Observaciones recientes sugieren que el universo es plano o muy próximo a plano.

2. La Cosmología Estándar

Modelos cosmológicos simples se estudian resolviendo las ecuaciones de Friedmann y de fluídos

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2}, \tag{10}$$

$$\dot{\rho} = -3H\left(\rho + p\right), \tag{11}$$

respectivamente, junto con la ecuación de estado $p = \omega\rho$. De aquí se sigue el comportamiento de los diversos componentes del universo: para un universo dominado por materia no-relativista: $\omega = 0, \rho \approx a^{-3}, a(t) \approx t^{2/3}, H = 2/3t$. En tanto que para uno dominado por materia relativista o radiación: $\omega = 1/3, \rho \approx a^{-4}, a(t) \approx t^{1/2}, H = 1/2t$. Si la componente dominante del universo es la energía del vacío tenemos que: $\omega = -1, \rho \approx \Lambda = constante, a(t) \approx e^{Ht}, H = cte$. Finalmente cuando se tiene una mezcla de componentes en donde ninguna domina (pero en donde todas contribuyen a la dinámica): $\rho = \rho_{mat} + \rho_{rad} + \rho_\Lambda + \cdots$. La densidad de número de partículas está dada por $n \approx a^{-3}$.

Lo anterior se modifica en el caso de curvatura distinta de cero. Usualmente la forma de proceder en los problemas sencillos es especificando el tipo de fluído que se estudia mediante su ecuación de estado. Posteriormente se resuelve la ecuación de fluído correspondiente Eq.(11) obteniéndose su desidad ρ como función del factor de escala y finalmente de la ecuación de Friedmann se sigue la evolución temporal del factor de escala. En el caso de un parámetro de estado ω constante las ecuaciones se pueden resolver en forma general. Para mezclas cada fluído obedecerá su propia ecuación de evolución, sin embargo, en cualquier caso, sólo habrá una ecuación de Friedmann.

Los principales parámetros que describen la dinámica global del universo son: la razón de expansión actual del universo H_0, las densidades (normalizadas) de energía Ω_i y el parámetro de desaceleración q_0. Las densidades Ω_i se definen como $\Omega_i \equiv \rho_i/\rho_c$ donde ρ_c se llama densidad crítica y se define como aquella requerida para hacer plana la geometría del universo.

Consideremos la edad del universo asociada con el parámetro de Hubble H_0. De la Ley de Hubble $v = H_0 r$, podemos ver fácilmente que H_0^{-1} tiene unidades de tiempo. Para un universo vacío $H_0 = 100h\frac{Km}{segMpc}$ implica

$$H_0^{-1} = 9{,}77h^{-1} \times 10^9 años, \tag{12}$$

donde, de acuerdo a las más recientes observaciones $h = 0{,}72 \pm 0{,}08$. A H_0^{-1} se le conoce como tiempo de Hubble. Para un universo plano dominado por materia la edad del universo se ve reducida con repecto al tiempo de Hubble dado que $a \approx t^{2/3}$, de donde se sigue que $H = 2/3t$ y por tanto $t_0 = (2/3)H_0^{-1}$. Para un universo abierto la edad aumentaría. Esto se entiende por el hecho de que en un universo con menos materia, toma más tiempo para que la atracción gravitacional detenga la expansión a la razón de expansión actual. Las observaciones sugieren que una mejor opción para un universo de baja densidad es mantener la geometría plana introduciendo una constante cosmológica (ρ_Λ se opone a la desaceleración). Los valores preferidos de los parámetros cosmológicos nos llevan a que $\Omega_{mat} \approx 0{,}3, h \approx 0{,}72, \rightarrow \Omega_\Lambda \approx 0{,}7, t \approx 13{,}7 \times 10^9$ *años*.

Como se menciona más arriba la densidad crítica se define como aquella requerida para hacer plana la geometría del universo y se denota ρ_c, queda definida por

$$\rho_c = 1{,}88h^2 \times 10^{-26}\frac{Kg}{m^3} = 2{,}78h^{-1} \times 10^{11}\frac{M_{sol}}{(h^{-1}Mpc)^3}, \tag{13}$$

donde M_{sol} denota una masa igual a la del sol. Una galaxia tiene masa $\approx 10^{11}M_{sol}$. Las galaxias típicamente están separadas $1Mpc$ de donde se sigue que la densidad del universo está muy próxima a la densidad crítica $\rho_{uni} \approx \rho_c$ y esto a su vez implica que $\Omega_{uni} \equiv \rho_{uni}/\rho_0 \approx 1$. La materia en las estrellas contribuye a la densidad de energía total con Ω_{estr} entre 0,005 y 0,01. No todo el material que podemos ver está en la forma de estrellas (también se observa gas, enanas obscuras con una masa $m \leq 0{,}08M_{sol}$, etc.). Nucleosíntesis implica la abundancia observada de elementos si $0{,}016 \leq \Omega_b h^2 \leq 0{,}024$. Por otro lado curvas de rotación de galaxias requieren un halo con $\Omega_{halo} \approx 0{,}1$. Asimismo la gravitación entre cúmulos de galaxias requiere que $\Omega_b \approx 0{,}35$, en tanto que por el movimiento peculiar de las galaxias se infiere $\Omega_b \geq 0{,}2$. Observaciones recientes apoyan una densidad total $\Omega_0 \approx 1$.

Todas estas son evidencias de que la materia bariónica conocida no puede explicar los diversos fenómenos observados y que claramente es necesaria la presencia de la materia obscura.

Mucho se ha avanzado en la comprensión del universo mediante el estudio de la radiación cósmica de fondo. El origen de la radiación de fondo se encuentra en la formación de sistemas neutros (átomos de hidrógeno) con la consecuente liberación de fotones a una temperatura $T_{dec} \approx 3000K$ conocida como época de recombinación o desacoplamiento. Esto ocurre en la superficie de última dispersión. Algunas propiedades importantes son su temperatura $T = 2{,}725 \pm 0{,}001K$ en la forma de radiación de cuerpo negro $\epsilon_{rad} = \rho_{rad}c^2 = \alpha T^4$ con una densidad de energía $\Omega_{rad} = 2{,}47 \times 10^{-5}h^{-2}$. Una propiedad importante se sigue de $\rho_{rad} \approx 1/a^4 \approx T^4$ y por tanto $T \approx 1/a$. Conforme el universo se expande y se enfría la distribución de radiación continúa correspondiendo a una distribución térmica de cuerpo negro.

Considerando la historia térmica del universo se tiene que la contribución a Ω_0 de neutrinos está dada por $\Omega_\nu \approx 1{,}68 \times 10^{-5}h^{-2}$. De manera que la contribución de partículas relativistas (neutrinos más fotones) es $\Omega_{rel} = \Omega_\gamma + \Omega_\nu \approx 4{,}15 \times 10^{-5}h^{-2}$ esto nos lleva a que la mayor parte de la materia del universo actual es no-relativista.

En la época de desacoplamiento $a_{dec} \approx 1/1000$ y por tanto $\Omega_{rel} < \Omega_{mat}$, de manera que durante el desacoplamiento el universo estaba dominado por materia no relativista. Para $\Omega_{rel} = \Omega_{mat}$ se sigue que para $a \equiv a_{eq} = 1/24000\Omega_0 h^2$ se da la época de igualdad de radiación y materia. Esto ocurre para $t_{eq} \approx 3400\Omega_0^{-3/2}h^{-3}años \approx 10,000\ años$, a una temperatura $T \approx 66,000\Omega_0 h^2 K$.

Nucleosíntesis estudia el origen de los elementos ligeros, ésta ocurre aproximadamente cuando $t \approx 1seg$, temperatura $T_{nucl} \approx 0{,}1Mev$ y el número de neutrones a protones es $N_n/N_p \approx 1/8$. La fracción de masa total de helio-4 es $Y \equiv 2N_n/(N_n + N_p) \approx 0{,}22$ o sea que 22 % de la materia del universo está en la forma de H_4, con hidrógeno $\approx 75\,\%$, deuterio $10^{-4}, H_3 \approx 10^{-5}$ y litio-7 $\approx 10^{-10}$.

3. Cosmología Inflacionaria

La cosmología hasta ahora descrita no está excenta de problemas. Éstos relacionados principalmente con condiciones iniciales o cuando se le considera en el contexto de teorías de gran unificación. Así es como, por ejemplo, tenemos el problema del universo plano. Esto es, dado que $|\Omega_{tot}(t) - 1| = |k|/a^2H^2$, para un universo dominado por radiación $a^2H^2 \approx t^{-1}$ y por tanto $|\Omega - 1| \approx t$ en tanto que para uno dominado por materia $a^2H^2 \approx t^{-2/3}$ y por tanto $|\Omega - 1| \approx t^{2/3}$. Se sigue que $|\Omega - 1|$ es una función creciente del tiempo. Esto implica que la geometría plana es una situación inestable para el universo. En otras palabras cabe preguntarse ¿cómo un universo que se ha estado desviando de un universo plano a lo largo de miles de millones de a*ñ*os se encuentra ahora tan aproximadamente plano? o equivalentemente, ¿porqué se encuentra tan cerca de la densidad crítica?

Otro problema de interés es el del horizonte: ¿cómo entendemos que regiones causalmente desconectadas de acuerdo a la cosmología del Big-Bang tengan sin embargo muy aproximadamente la misma temperatura? Esto se infiere de las mediciones de la radiación de fondo.

Teorías de gran unificación en conjunción con cosmología predicen la existencia de abundantes defectos topológicos (monopolos magnéticos, cuerdas cósmicas, etc.) que sin embargo no se observan. Más importante es el hecho de que en la cosmología del Big-Bang las condiciones iniciales necesarias para la formación de estructura observada (cúmulos de galaxias y supercúmulos) se introducen a "mano" y no son obtenidas de la teoría.

Inflación ofrece una explicación plausible para la resolución de todos los problemas mencionados arriba. En su forma más general inflación se define como una época de expansión del universo en que el factor de escala crece en forma acelerada, esto es inflación ocurre cuando $\ddot{a} > 0$,

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p), \quad p < -\frac{\rho}{3} \quad \rightarrow \quad \ddot{a} > 0. \tag{14}$$

Por ejemplo, para una constante cosmológica la ecuación de estado es $p = -\rho$. Según la ecuación de Friedmann

$$H^2 = \frac{8\pi G}{3}\rho - \frac{k}{a^2} + \frac{\Lambda}{3}, \tag{15}$$

después de un tiempo Λ domina y por tanto $H^2 \approx \frac{\Lambda}{3}$ de donde se sigue que $a(t) = e^{(\frac{\Lambda}{3})^{1/2}t}$. Este crecimiento exponencial permite resolver los problemas de la cosmología del Big-Bang. En general sólo se requiere que el factor de escala crezca cuasi-exponencialmente pero siempre de tal manera que $\ddot{a} > 0$.

El problema importante de la formación de estructura se entiende ahora de la manera siguiente: inflación usualmente se implementa con el auxilio de la física de partículas elementales. En los modelos de partículas existen campos escalares. Un campo escalar, a través de su energía potencial, puede dar lugar a una época inflacionaria. Este campo escalar está (como cualquier otro) sujeto a fluctuaciones cuánticas las cuales son amplificadas hasta escalas macroscópicas por el proceso inflacionario. La evolución de estas fluctuaciones descritas por las ecuaciones linealizadas de Einstein son las que eventualmente darán lugar (por inestabilidad gravitacional) a la formación de la estructura conocida.

La cosmología inflacionaria constituye otro valioso ejemplo en la mejor tradición científica donde diversas ramas de la física se suman para comprender y describir mejor los fenómenos naturales. En el caso de la cosmología, cuando la teoría de la Relatividad General se aplica

al universo ésta se constituye en el pilar teórico más sólido en la descripción de la dinámica del universo. Otras etapas cruciales en la evolución del universo tales como nucleosíntesis y recombinación requieren de la física nuclear y atómica respectivamente para entender los procesos microscópicos que se dieron en el universo y que generaron las abundancias de elementos ligeros primordiales así como la radiación cósmica de fondo. De manera análoga la cosmología inflacionaria típicamente usa campos escalares para producir inflación a través de su energía potencial. Estos campos son comunes en teorías microscópicas de la materia tales como teorías de gran unificación y sus extensiones supersimétricas, así como en teorías de cuerdas y branas que son extensiones de los modelos de partículas y que incorporan a la gravedad en su descripción.

Agradezco al Dr. José Récamier por la gentileza de invitarme a participar en la XIX Escuela de Verano en Física, mediante la impartición de una plática, resumida en estas notas.

Referencias

[1] E.W. Kolb and M.S. Turner, *The Early Universe*, Addison-Wesley, Pub. Co., 1990.

[2] P.J.E. Peebles, *Principles of Physical Cosmology*, Princeton U. Press, 1993.

[3] J.A. Peacock, *Cosmological Physics*, Cambridge U. Press, 1999.

[4] A. Liddle, *An Introduction to Modern Cosmology*, Sec Ed., John Wiley & Sons, 2003.

[5] S. Dodelson, *Modern Cosmology*, Academic Press, 2003.

[6] S. Weinberg, *Cosmology*, Oxford University Press, 2008.

Métodos en Física Molecular para la detección de trazas moleculares y sus aplicaciones

Notas de la escuela de Verano, EVER2012

Dr. Antonio Marcelo Juárez Reyes

Instituto de Ciencias Físicas UNAM *juarez@fis.unam.mx*

CONTENIDO

1.- Definiciones y conceptos generales

En este documento se presenta una visión general e introductoria del estudio de trazas moleculares, las técnicas que se emplean para detectarlas y las aplicaciones de esta área emergente de la física. Se incluyen algunas referencias útiles para que el estudiante interesado pueda, si así lo deseare, profundizar en el tema. Asimismo, este capítulo describe, de manera general, las técnicas más comunes y mejor establecidas en el el tema de detección de trazas moleculares, con el fin de enterar al alumno interesado en la existencia de estas técnicas Como documento introductorio y necesariamente general, no se presentan detalles específicos de estas técnicas, en aras de resaltar las aplicaciones de esta área de la física molecular. El alumno interesado en profundizar el tema puede consultar una revisión en resúmenes especializados [**1**]

Se definen las trazas moleculares en el presente trabajo como aquellas especies moleculares que se encuentran diluidas en fase gaseosa en nuestra atmósfera, en partes de 1 por mil millones volumétrica o menores. Estas moléculas ultra-diluídas puden tener su origen en distintas fuentes, por ejemplo, pueden corresponder a moléculas contaminantes en la atmósfera [**2**], a marcadores metabólicos generados en un proceso

biológico o biomédico específico [**3**], a hormonas de comunicación entre plantas [**4**], entre otras posibilidades. A este nivel de dilución, las técnicas convencionales de espectroscopia de absorción, emisión o químicas no tienen la capacidad de cuantificar la presencia de un tipo particular de molécula diluida de manera precisa. Es justamente la cuantificación de las trazas moleculares, además de su identificación lo que hace tan relevante este campo [**5**].

Para ilustrar la utilidad de la técnica de detección de trazas moleculares mencionaré varios ejemplos a continuación de áreas o usos en los que la capacidad de detectar y cuantificar la presencia de moléculas en la atmósfera es relevante en casos prácticos.

a.- Control de tiempos de maduración de frutas (Post harvest control) Algunas plantas y frutos sincronizan sus procesos y se □comunican□ por medio de la emisi□n de etileno[**4**]. Sin embargo, para poder estudiar (y eventualmente controlar) estos procesos de comunicación se debe ser capaz de detectar en la atmósfera niveles de etileno que lo califican como una traza molecular. La comunicación entre plantas en un invernadero puede alcanzar niveles de decenas de partes por mil millones (ppmm). El desarrollo de sensores de etileno a estos niveles y la implementación de sistemas de lazo cerrado para regular el nivel de etileno en bodegas o camiones de transporte de frutas es actualmente un reto tecnológico que se está implementando a nivel comercial en Europa, principalmente, donde el transporte de frutos y el correcto control de sus tiempos de maduración es fundamental [**6**].

b.- Detección temprana de fugas o acumulación de gases explosivos en minas. Cada año a nivel mundial mueren cientos de mineros por accidentes relacionados con la acumulación de gas grisú. Nuestro país no es ajeno a este tipo de eventos tan desafortunados. Un sistema de detección de gas grisú, dotado de la ingeniería necesaria para generar alertas y reportar de manera automatizada los niveles de gases explosivos en una mina es de indudable importancia y las técnicas de detección de trazas moleculares proporcionan una alternativa eficiente, económicamente viable y técnicamente asequible.

c.- Detección de substancias ilícitas o explosivos [**7**]. Los sistemas de seguridad de los aeropuertos en nuestro país dependen totalmente, en la actualidad, de sistemas de detección de explosivos y substancias ilícitas basados en detectores de trazas moleculares, típicamente del tipo de tubo de deriva iónica.

d.- Método de diagnóstico médico por análisis no invasivo del aliento humano. Esta técnica, que tiene como principio cuantificar la presencia de marcadores de enfermedades (la acetona, por ejemplo, en rangos de 10 partes por millón en el aliento humano indica la presencia de diabetes *mellitus* del tipo 2) se encuentra actualmente en proceso de desarrollo en Europa y en Estados Unidos [**8**]. El potencial económico, social y humano que una técnica de diagnóstico de este tipo tiene sobre el bienestar de la población es muy grande. No hay técnica menos invasiva que la del análisis del aliento humano. La base de esta técnica tan importante es, esencialmente, la detección de trazas moleculares en el aliento humano en partes por mil millones.

e.- Control de calidad de café. La industria del café, en términos económicos, es casi tan grande como la del petróleo. Nuestro país es un productor importante de este producto y los procesos de control de calidad, evaluación del tipo de café y la selección de este es un proceso con un valor comercial muy elevado. Una máquina o proceso capaz de informar a un cliente potencial de la calidad de un lote particular de café tiene un valor comercial muy grande. Las técnicas de detección de trazas moleculares, aunadas a métodos computacionales de análisis de componentes (tales como los algoritmos genéticos y el algoritmo PCA □principal component analysis-) permiten en la actualidad realizar estudios diferenciadores de la calidad y el tipo de un grano particular, a partir del análisis de las trazas moleculares generadas en el proceso de tostado [**9**].

f.- Prospección petrolera. A grandes rasgos, la prospección de una zona petrolera con potencial económico es un proyecto económicamente muy caro (cientos de millones de dólares) y complejo. Sin embargo, las zonas de riqueza petrolera potencial pueden detectarse de manera preliminar por un fen□meno conocido como □Microseepage an□malo□en t□rminos sencillos, el filtrado de trazas moleculares (radón, butano y otros hidrocarburos ligeros) a través de los anticlinales que sellan a una reserva subterránea □marcan□ la frontera en donde hay un pozo potencialmente □til de petr□leo. Usando técnicas de detección de las trazas moleculares arriba mencionadas, se puede realizar una prospección preliminar relativamente barata y eficiente [**10**].

Por razones de espacio, y dado el carácter genera de este resumen se omite dar más detalles, pero para dar un panorama general, podemos mencionar otros ejemplos: el estudio de tasas metabólicas en el cuerpo humano o en animales usando medicamentos marcados isotópicamente y monitoreando la exhalación de CO_2 marcado isotópicamente en el aliento [**11**]. Asimismo, estudios de el ciclo del carbono estudiando trazas moleculares en hielos polares, estudios de c□lulas senescentes al analizar los □vapores□ exhalados por estas células en una caja Petri donde se les esté cultivando, así como mediciones de la emisión de vapores tóxicos en adhesivos usados en usos domésticos, con el fin de determinar que se encuentran en la norma mexicana y no será causa de enfermedades crónicas al largo plazo. Cabe mencionar que todos los ejemplos mencionados arriba son relevantes en la economía de nuestro país y el bienestar de sus ciudadanos que son, a final de cuentas, quienes pagan la cuenta de la investigación y desarrollo que se hace en nuestras universidades, y a quienes debemos nuestro quehacer como científicos o estudiantes. Para concluir esta sección, y en resumen, la detección de trazas moleculares en partes por mil millones o partes por millón es un tema relevante en varias áreas de la ciencia, tanto fundamental como aplicada.

¿cómo se implementa este tipo de técnicas?

Cada una de estas técnicas viene en varias presentaciones y modalidades. Una parte común que las une, es que estas técnicas requieren de la sinergia de la física molecular avanzada, con instrumentación en ingeniería de alto nivel Asimismo, todas estas

técnicas comparten la necesidad de electrónica analógica y digital avanzadas, combinadas con técnicas de vacío, espectroscopia molecular avanzada, algoritmos de álgebra lineal y el conocimiento especializado que cada una de las ramas en donde tiene aplicación (Medicina, biología, defensa y seguridad nacional, agricultura protegida). En ese sentido, el área de detección de trazas moleculares es muy rica y necesariamente multidisciplinaria. A continuación presentaré, de maneara muy general, los principios de operación de cada una de estas técnicas. Estos instrumentos y técnicas se encuentran a disposición del grupo de Física Atómica, Molecular y Óptica del Instituto de Ciencias Físicas y su red de colaboradores. Los estudiantes interesados en aprender, trabajar o aplicar alguna o varias de estas técnicas puede visitarnos y formar parte del equio de desarrollo en alguno de los campos de aplicación y uso que hemos mencionado arriba.

Para darle a los estudiantes una idea de las técnicas disponibles actualmente en México para este tipo de estudios, relacionados con la detección de trazas moleculares en fase gaseosa.

2 **Técnicas de cavidades**

La técnica de cavidades ópticas hace uso de resonadores ópticos (un par de espejos parabólicos, de alta reflectividad colocados frente a frente) para amplificar la absorción de la luz láser dentro de la cavidad (BBCEAS, broad band cavity enhanced absorption spectroscopy) o bien para medir el tiempo medio de vida de un fotón dentro de la cavidad (Cavity ring down spectroscopy [12]). El arreglo experimental típico de Cavity Ring Down consiste en un laser que se emplea para iluminar una cavidad óptica de alta "fineza", es decir, que est compuesta de espejos con una reflectividad muy cercana a 1 (en el caso de la cavidad del ICF la reflectividad de nuestros espejos es del 99.998%). Suponiendo que en la cavidad óptica no hay ninguna especie molecular presente, ocurre lo siguiente: Cuando la frecuencia del laser se encuentra en resonancia con uno de los modos de la cavidad, la luz se acumula dentro de la cavidad, aunque parte de esta escapa, debido a que la reflectividad de los espejos no es exactamente igual a 1. Si el laser es pulsado, la intensidad de salida obedece a una ley exponencial. La constante de decaimiento depende del valor de la reflectividad y se puede medir de manera sencilla con un osciloscopio. Durante este decaimiento, la luz va y viene de un espejo a otro muchas veces, de tal suerte que la trayectoria óptica de el rayo de luz es del orden de miles o decenas de miles de metros de longitud efectiva. Si uno coloca una muestra molecular dentro de la cavidad, tal que esta muestra tenga una transición resonante con la frecuencia del laser, esta especie absorberá la luz. En este caso, la constant original de decaimiento exponencial se modificará (existe manera de cuantificar cuanto) de tal manera que, midiendo la constant de decaimineto con la muestra molecular presente, y conmparándola con aquella medida con la cavidad vacía, se puede cuantificar cuántas moléculas se encuentran presentes por unidad de volumen en la cavidad. El nivel de sensitividad de esta técnica alcanza valores de una parte por un millon de millones de moléculas, en una unidad de volume dada.. Hay otros sabores de estas t cnias, tales como el "cavity leak out spectroscopy" y las t cnicas de "multipass cells" que emplean espejos acrom ticos para magnificar la trayectoria de un haz de luz dentro de una cavidad. Todas ellas se basan en el uso de fuentes de luz infraroja, espejos de alta reflectividad y tienen sensitividades muy elevadas. La figura 1 muestra, de manera esquemática, el arreglo experimental necesario para implementar la técnica de CRD. El equipo es relativamente sencillo y no excesivamente caro, salvo por

la fuente de luz, por lo que puede implementarse de manera sencilla. En el Instituto de Ciencias Físcas de la UNAM contamos actualmente con un sistema de este tipo.

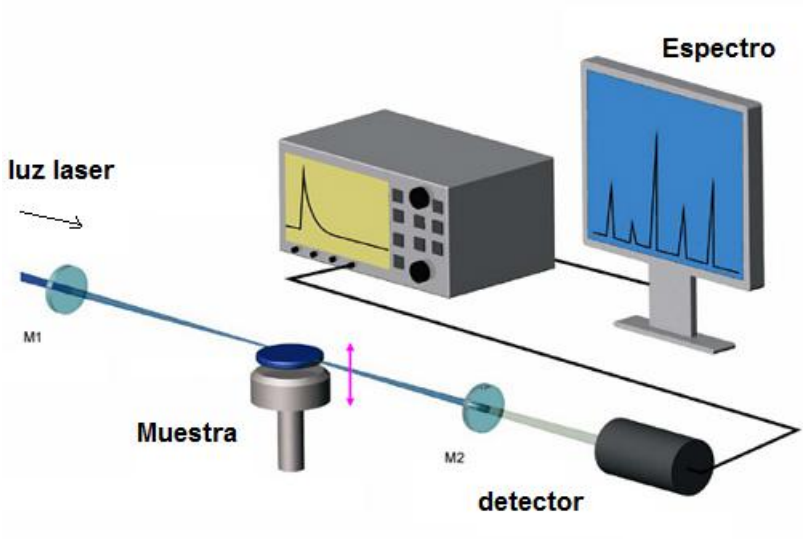


Figura 1. Arreglo experimental básico de Cavity Ring Down. Otras formas de medición con cavidades tienen arreglos semejantes, auque los principios de operación son diferentes.

3 **Técnicas de espectroscopía por transferencia de protones**

La espectroscopía por transferencia de protones (proton transfer spectroscopy) es una de las técnicas más poderosas y sensibles para detectar moléculas diluídas en la atmósfera[**13**]. El principio físico de este instrumento se basa en el proceso de transferencia de carga entre una molécula protonada de agua H_3O^+ y el compuesto a analizar, M

$$H_3O^+ + M \longrightarrow MH^+ + H_2O$$

Se puede probar que el número de procesos de transferencia de protones como función del tiempo [MH+]t es igual a

$$[MH^+]_t = [H_3O^+]\, k\, [M]\, t$$

Donde

$[H_3O^+]$ = numero original de iones hidronio

$[MH^+]_t$ = número de iones M creados por unidad de tiempo

k = Constante de reacción

[M] = densidad de la muestra

El esquema experimental con el que se logra esto se muestra en la figura 2

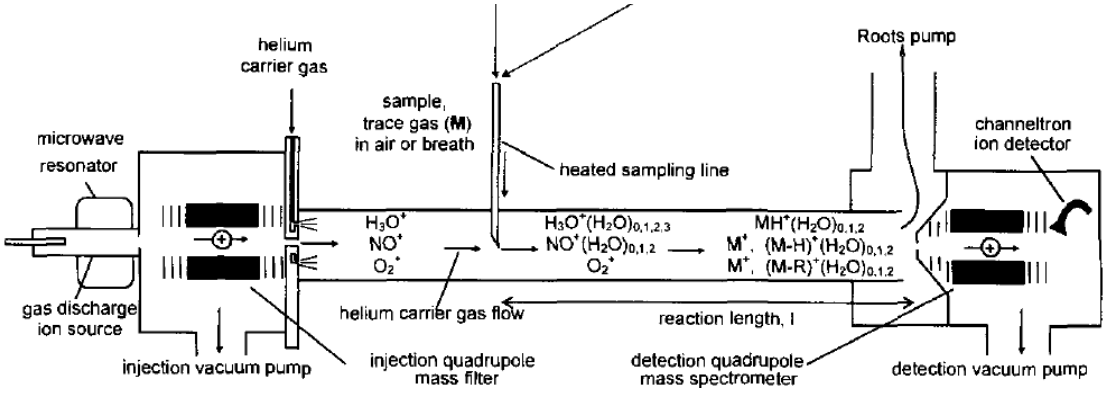


Figura 2. Esquema experimental de un espectrómetro por transferencia de protones. EN el ICF contamos con un tubo de deriva que puede emplearse para estos fines.

Esta técnica es muy poderosa y se usa actualmente en usos que van desde la física médica (análisis de aliento exhalado) a la cata de vinos y café de manera electrónica(referencia, nariz electrónica Austria).

4.- Técnicas de cromatografía de gases con análisis de masas e ionización

La cromatografía de gases basa su principio en la movilidad diferencial de moléculas al fluir a travez de un capilar. Las moléculas más pequeñas y menos polares se desplazan con mayor rapidez que las moléculas grandes y polares. A la salida del cromatógrafo se analiza el tiempo de salida de las especies, que está relacionado a la especie molecular específica. Una segunda etapa de ionización, que puede ser una trampa de i9ones, o bien un tubo de tiempo de vuelo, y análisis de masas permite separar con mayor sensitividad a las moléculas [**14**].

Existen varios tipos de cromatografía de gases (GC): la cromatografía gas-sólido (GSC), en fase líquida (HPLC) y la cromatografía gas-líquido (GLC), siendo esta última la que se utiliza más ampliamente, y que se puede llamar simplemente cromatografía de gases (GC). En la GSC la fase estacionaria es sólida y la retención de los analitos en ella se produce mediante el proceso de adsorción. Precisamente este proceso de adsorción, que no es lineal, es el que ha provocado que este tipo de cromatografía tenga aplicación limitada, ya que la retención del analito sobre la superficie es semipermanente y se obtienen picos de elución con colas. Su única aplicación es la separación de especies gaseosas de bajo peso molecular. La GLC utiliza como fase estacionaria moléculas de líquido inmovilizadas sobre la superficie de un sólido. La figura 3 muestra, de manera esquemática, el arreglo experimental empleado en la implementación de la cromatrografía de gases,sea esta gaseosa, líquida o en fase

gaseosa. inerte.

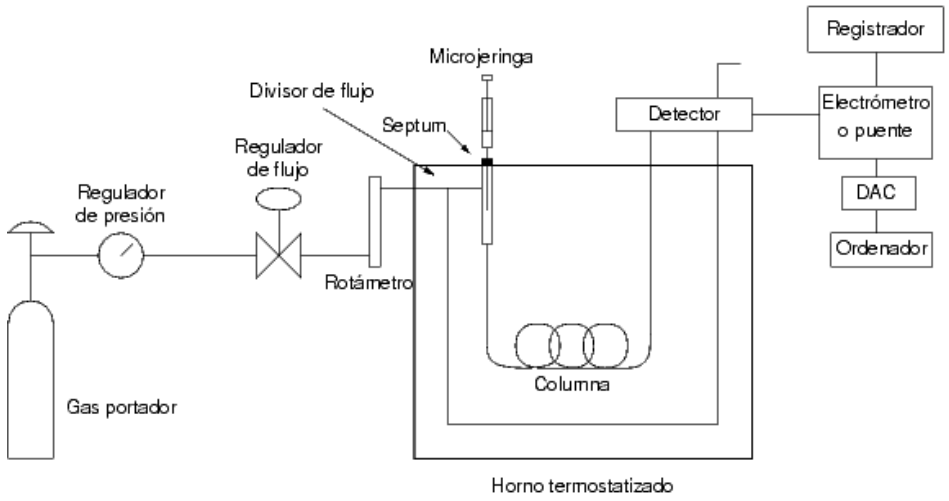


Figura 3. Arreglo experimental general emleado, en el caso de cromatografía en fase liquida, en un cromatr□grafo de gases. El t□rmino □cromat□grafo □proviene del hecho histórico de que este tipo de instrumentos, se empleaban para separar tintas de distointos colores.

5 Tecnicas de Laser Induced flourescence y FTIR

La técnica de fluorescencia inducida por laser se basa en la excitación de la muestra a analizar, seguida del estudio de la fluorescencia emitida. Esta fluorescencia se puede medir de manera muy sensible usando técnicas de conteo fotónico. La longitu de onda de excitación se elije de tal manera que la sección transversal de excitación sea la más grande posible. Las aplicaciones prácticas de esta técnica son muy diversas, aunque el costo de un instrumento de LIF es relativamente caro. La figura 4 muestra el esquema experimental necesario para esta técnica

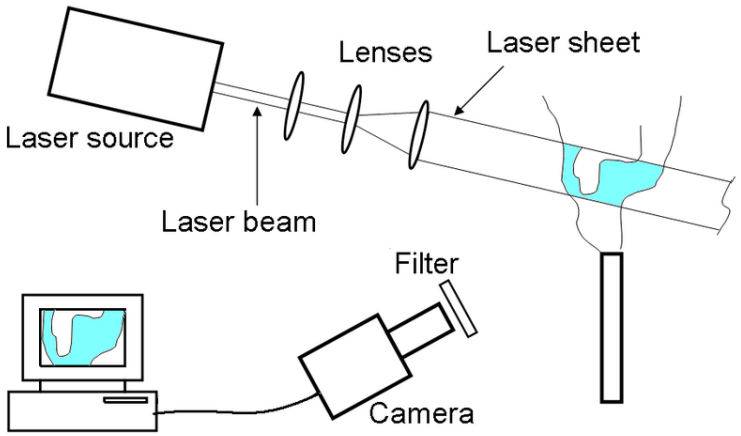


Figura 4. Arreglo exérimental para llevar a cabo espectroscopía molecular por fluorescencia inducida por laser.

FTIR (Fourier transformed Infrared spectroscopy) es una técnica para obtener espectros de absorción de moléculas basada en el uso de la transformada inversa de fourier de un interferograma [**15**]. Un espectrómetro FTIR consta de una fuente infrarrona de ancho de banda grande, un interferómetro del tipo de Michelson y un espejo móvil (ver figura 3). Al mover el espejo se generan interferogramas que cambian con la posición de el brazo.Al aplicar la transformada inversa de Fourier a este interferograma se obtiene el espectro de absorción de las moléculas bajo estudio. La ventaja de FTIR es que permite realizar estudios de fases sólidas, gaseosas o líquidas.

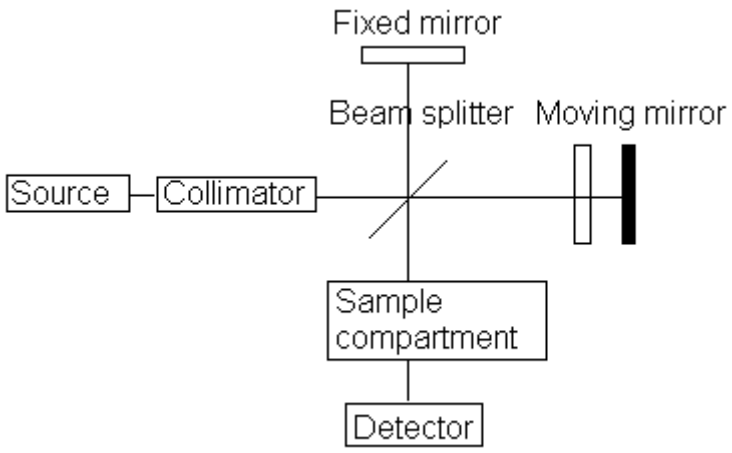


Figura 3. Arreglo experimental típico de un espectrómetro FTIR.

6.- Estudio de trazas en México y su futuro.

Dada la relevancia en tantas áreas de investigación, tanto fundamentales como aplicadas, los estudios basados en técnias de cavidades son muy útiles y poderosos. En este sentido, en el instituto de Ciencias Físicas de la UNAM estamos desarrollando actualmente un grupo especializado en el estudio de trazas moleculares y su aplicación en distintas áreas. Para este propósito, nuestro grupo está incorporando y desarrollando la infraestructura y las redes de colaboración necesarias para contar con un laboratorio avanzado de detección molecular. En este sentido, actualmente nuestra red cuenta, en un solo lugar, con la infraestructura necesaria para llevar a cabo estudios de Proton Transfer Spectroscopy, Cavity Ring Down y otras variantes. Sumando a esto la colaboración con la facultad de Ciencias de la UNAM y el departamento de química de la Universidad Autónoma de Morelos, así como el grupo de óptica y Física molecular aplicada de la Universidad de Delft, en conjunto, nuestro grupo cubre todas las áreas necesarias, en cuanto a infraestructura y capacidad técnica, mencionadas arriba. La segunda etapa de este interesante y ambicioso proyecto, consiste en usar esta infraestructura para atender áreas de la biología, la medicina y el medio ambiente, relevantes para nuestro país. Mientras escribo estas notas, nuestro grupo se encuentra en una etapa muy interesante de desarrollo y, hasta donde alcanza nuestro mejor conocimiento, no existe en México ni en Latinoamérica un esfuerzo semejante. Laboratorios de este tipo, al nivel de infraestructura con el que contamos en la UNAM

sólo se puede encontrar actualmente en Delft, Holanda y Duserdolf en Alemania, grupos con los que nuestro laboratorio tiene ligas actualmente. Este momento es muy interesante en nuestro laboratorio, y cierro este documento invitando a los estudiantes interesados en realizar estudios avanzados de física molecular de alto nivel, e interesados en que estos estudios repercutan en la solución de problemas reales y relevantes del país, a contactarnos y ser parte de nuestro equipo de trabajo.

Cuernavaca Morelos, 26 de enero del 2011.

Referencias

[**1**] Review of Scientific Instruments, 75, 8, page 2499 (2004)

[**2**] Atmospheric Environment Volume 34, Issues 12–14, 2000, Pages 2063–2101

[**3**] IEEE Sensors Journal Vol. 10, NO. 1, 2010

[**4**] Environ. Microbiol. November 2002 vol. 68 no. 11 5342-5350

[**5**] Rev. Sci Instrum. 79, 123110 (2008)

[**6**] Annals of applied biology, 1985

[**7**] Rev. Sci. Instrum. **75**, 2499 (2004)

[**8**] *Anal. Chem.*, **2010**, *82* (9), pp 3581–3587

[**9**] Applied Spectroscopy, Vol. 49, Issue 5, pp. 580-585 (1995)

[**10**] Remote Sensing Reviews Volume 18, Issue 1, 2000

[**11**] Mini Reviews in Medicinal Chemistry, Volume 7, Number 2, February 2007 , pp. 115-129(15)

[**12**] http://www.laserfocusworld.com/articles/print/volume-37/issue-5/features/spectroscopy/cavity-ringdown-technique-measures-absorption.html

[**13**] *J. Am. Chem. Soc.*, **1983**, *105* (15), pp 5133–5134

[**14**] http://www.sigmaaldrich.com/etc/medialib/docs/Aldrich/Bulletin/1/the-basics-of-gc.Par.0001.File.tmp/the-basics-of-gc.pdf

[**15**] International Biodeterioration & Biodegradation Volume 41, Issue 1, 1998, Pages 1–11

Momento Angular en Bases de Armónicos Esféricos y Esferoconales

Eugenio Ley Koo, Instituto de Física, UNAM
eleykoo@fisica.unam.mx

October 24, 2011

1 Introducción

El tema de esta conferencia está asociado a trabajos de investigación reportados en la publicación "Rotaciones de Moléculas Asimétricas y el Átomo de Hidrógeno en Configuraciones Libres y Confinadas"[1], especialmente en su sección 4 "Sobre el Desarrollo de la Teoría de Momento Angular en Base de Armónicos Esferoconales de Lamé".

La Conferencia misma y esta versión escrita para las Memorias de la XIX Escuela de Verano en Física tienen el propósito de servir a alumnos y lectores como introducción al tema específico, partiendo de sus conocimientos previos y familiaridad con el tema en coordenadas cartesianas y esféricas. Las secciones subsecuentes están dedicadas sucesivamente a: 2) identificar los operadores de cantidad de movimiento lineal, momento angular, Laplaciano y Hamiltoniano para rotaciones de moléculas asimétricas; 3) construir las funciones armónicas cartesianas, esféricas y esferoconales para $\ell = 0, 1, 2, 3$ como ilustración; y 4) mencionar algunas aplicaciones en investigaciones en curso.

2 Operadores de gradiente, divergencia, Laplaciano, momento lineal, momento angular y Hamiltoniano de molé culas asimétricas.

Los operadores de gradiente, divergencia y Laplaciano son conocidos desde el curso introductorio de electricidad y magnetismo. Efectivamente, el primero establece la conexión entre el potencial electrostático $\phi(\overrightarrow{r})$, función escalar, y el campo vectorial de intensidad eléctrica $\overrightarrow{E}(\overrightarrow{r})$, en la forma,

$$\overrightarrow{E}(\overrightarrow{r}) = -\operatorname{grad}\phi(\overrightarrow{r}) \equiv -\bigtriangledown\,\phi(\overrightarrow{r}). \tag{1}$$

El segundo aplicado a la intensidad de campo eléctrico conduce a la densidad volumétrica de carga eléctrica $\rho(\overrightarrow{r})$, según la ley de Gauss:

$$\operatorname{div}\overrightarrow{E} \equiv \nabla \cdot \overrightarrow{E}(\overrightarrow{r}) = 4\pi\rho(\overrightarrow{r}). \tag{2}$$

El operador Laplaciano definido como la divergencia del gradiente aparece al sustituir la Ec. (1) en la Ec. (2):

$$-\operatorname{div}(\operatorname{grad}\phi\overrightarrow{r}) = -\nabla\cdot\nabla\phi \equiv -\nabla^2\phi = 4\pi\rho \tag{3}$$

conocida como la ecuación de Poisson.

En coordenadas cartesianas el operador nabla,

$$\nabla = \widehat{i}\frac{\partial}{\partial x} + \widehat{j}\frac{\partial}{\partial y} + \widehat{k}\frac{\partial}{\partial z} \tag{4}$$

permite escribir las ecuaciones respectivas en las formas

$$\overrightarrow{E}\,(x,y,z) = -\left(\widehat{i}\frac{\partial\phi}{\partial x} + \widehat{j}\frac{\partial\phi}{\partial y} + \widehat{k}\frac{\partial\phi}{\partial z}\right) \tag{5}$$

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 4\pi\rho\,(x,y.z) \tag{6}$$

$$-\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\phi\,(x,y.z) = 4\pi\rho\,(x,y.z) \tag{7}$$

En aquellos puntos donde la densidad de carga eléctrica es nula, la ecuación de Poisson se reduce a la llamada ecuación de Laplace,

$$\nabla^2\phi\,(\overrightarrow{r}) = 0. \tag{8}$$

Las soluciones de la ecuación de Laplace se conocen con el nombre de funciones armónicas. En la sección 3 se ilustra la construcción sistemática de tales soluciones en coordenadas cartesianas, esféricas y esferoconales.

Consideremos primero las ecuaciones de transformación entre las coordenadas respectiva $(x,y.z)$, (r,θ,φ) y (r,χ_1,χ_2) :

$$\begin{aligned} x &= r\,\mathrm{sen}\,\theta\cos\varphi = r\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2) \\ y &= r\,\mathrm{sen}\,\theta\cos\varphi = r\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2) \\ z &= r\cos\theta = r\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_2|k_2)\,. \end{aligned} \tag{9}$$

Las dos últimas comparten la coordenada radial $r = \sqrt{x^2+y^2+z^2}$ que define esferas concéntricas en el origen. Mientras las coordenadas esféricas familiares definen direcciones

en el espacio en términos del ángulo polar $0 \leqslant \theta \leqslant \pi$, que define conos circulares con el eje z como el eje polar común, y del ángulo azimutal $0 \leqslant \varphi \leqslant 2\pi$, que define semi-planos meridianos que giran alrededor del eje z ; las coordenadas esferoconales están definidas en términos de funciones elípticas de Jacobi [2]:

$$u = \int_0^{\beta} \frac{d\alpha}{\sqrt{1 - k^2 \operatorname{sen}^2 \alpha}} \tag{10}$$

de modo que

$$\begin{aligned} \operatorname{sn}(u|k) &= \operatorname{sen}\beta \\ \operatorname{cn}^2(u|k) &= \cos^2\beta = 1 - \operatorname{sn}^2\beta \\ \operatorname{dn}^2(u|k) &= 1 - k^2 \operatorname{sn}^2\beta \end{aligned} \tag{11}$$

donde β se llama la amplitud de u. Las derivadas con respecto a su argumento u son

$$\begin{aligned} \frac{d}{du}\operatorname{sn}(u|k) &= \operatorname{cn}(u|k)\operatorname{dn}(u|k) \\ \frac{d}{du}\operatorname{cn}(u|k) &= -\operatorname{sn}(u|k)\operatorname{dn}(u|k) \\ \frac{d}{du}\operatorname{dn}(u|k) &= -k^2 \operatorname{sn}(u|k)\operatorname{cn}(u|k) \end{aligned} \tag{12}$$

involucrando al producto de las otras dos funciones. Los parámetros k_1 y k_2 en las funciones elípticas de las Ecs. (9) están restringidos por la condición $k_1^2 + k_2^2 = 1$. El lector puede analizar las situaciones límite cuando $k \to 0$, para la cual $u = \beta$ y $\operatorname{dn}(u|0) = 1$; y cuando $k \to 1$, para la cual $\operatorname{dn}(u|1) = \operatorname{cn}(u|1)$; en ambos casos las coordenadas esferoconales se transforman en coordenadas esféricas. También puede usar las Ecs. (9) y (11) para establecer que

$$\begin{aligned} \frac{x^2}{\operatorname{dn}^2(\chi_1|k_1)} + \frac{y^2}{\operatorname{cn}^2(\chi_1|k_1)} &= r^2 = x^2 + y^2 + z^2 \\ \frac{y^2}{\operatorname{cn}^2(\chi_2|k_2)} + \frac{y^2}{\operatorname{dn}^2(\chi_2|k_2)} &= r^2 = x^2 + y^2 + z^2 \end{aligned} \tag{13}$$

y por lo tanto:

$$\begin{aligned} \frac{x^2}{z^2 \frac{\operatorname{dn}^2(\chi_1|k_1)}{k_1^2 \operatorname{sn}^2(\chi_1/k_1)}} + \frac{y^2}{z^2 \frac{\operatorname{cn}^2(\chi_1|k_1)}{\operatorname{sn}^2(\chi_1|k_1)}} &= 1 \\ \frac{y^2}{x^2 \frac{\operatorname{cn}^2(\chi_2|k_2)}{\operatorname{sn}^2(\chi_2|k_2)}} + \frac{z^2}{x^2 \frac{\operatorname{dn}^2(\chi_2|k_2)}{k_2^2 \operatorname{sn}^2(\chi_2|k_2)}} &= 1 \end{aligned} \tag{14}$$

Las dos últimas ecuaciones permiten reconocer que los lugares geométricos definidos por valores fijos de $\chi_1 = \chi_{10}$ y $\chi_2 = \chi_{20}$ corresponden respectivamente a conos con eje a lo largo del eje z y secciones elípticas con semieje mayor $z\,\mathrm{dn}\,(\chi_{10}|k_1)\,/k_1\,\mathrm{sn}\,(\chi_{10}|k_1)$ paralelo al eje x y semieje menor $z\,\mathrm{cn}\,(\chi_{10}|k_1)\,/\,\mathrm{sn}\,(\chi_{10}|k_1)$ paralelo al eje y, y con eje a lo largo del eje x y secciones elípticas con semieje mayor $x\,\mathrm{dn}\,(\chi_{20}|k_2)\,|k_2\,\mathrm{sn}\,(\chi_{20}|k_2)$ paralelo al eje z y semieje menor $x\,\mathrm{cn}\,(\chi_{20}|k_2)\,|\,\mathrm{sn}\,(\chi_{20}|k_2)$ paralelo al eje y.

Los conjuntos de estos conos elípticos (χ_{10}, χ_{20}) definen las posiciones sobre las esferas de radio r. Para valores fijos de r, las Ecs. (13) describen las intersecciones de los conos elípticos respectivos con la esfera correspondiente.

A continuación se evalúan los desplazamientos espaciales en las diferentes coordenadas:

$$
\begin{aligned}
d\overrightarrow{r} &= \widehat{i}dx + \widehat{j}dy + \widehat{k}dz \\
&= \left(\widehat{i}\,\mathrm{sen}\,\theta\cos\varphi + \widehat{j}\,\mathrm{sen}\,\theta\,\mathrm{sen}\,\varphi + \widehat{k}\cos\theta\right)dr \\
&\quad + r\left[\cos\theta\left(\widehat{i}\,\mathrm{sen}\,\varphi + \widehat{j}\,\mathrm{sen}\,\varphi\right) - \widehat{k}\,\mathrm{sen}\,\theta\right]d\theta \\
&\quad + r\,\mathrm{sen}\,\theta\left(-\widehat{i}\,\mathrm{sen}\,\varphi + \widehat{j}\cos\varphi\right)d\varphi \\
&= \left[\widehat{i}\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2) + \widehat{j}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2) + \widehat{k}\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_2|k_2)\right]dr \\
&\quad + r\left[-\widehat{i}k_1^2\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2) - \widehat{j}\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2)\right. \\
&\quad \left. + \widehat{k}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_2|k_2)\right]d\chi_1 \\
&\quad + r\left[\widehat{i}\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2)\,\mathrm{dn}\,(\chi_2|k_2) - \widehat{j}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)\,\mathrm{dn}\,(\chi_2|k_2)\right. \\
&\quad \left. - \widehat{k}k_2^2\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)\,\mathrm{cn}\,(\chi_2|k_2)\right]d\chi_2 \\
&= \widehat{r}h_r dr + \widehat{\theta}h_\theta d\theta + \widehat{\varphi}h_\varphi d\varphi = \widehat{r}h_r dr + \widehat{\chi}_1 h_{\chi_1} d\chi_1 + \widehat{\chi}_2 h_{\chi_2} d\chi_2. \qquad (15)
\end{aligned}
$$

En el último renglón se introducen los vectores unitarios y factores de escala asociados a las respectivas coordenadas, y que se identifican a partir de los renglones respectivos más arriba:

$$
\begin{aligned}
\widehat{r} &= \widehat{i}\,\mathrm{sen}\,\theta\cos\varphi + \widehat{j}\,\mathrm{sen}\,\theta\,\mathrm{sen}\,\varphi + \widehat{k}\cos\theta \\
\widehat{\theta} &= \mathrm{sen}\,\theta\left(\widehat{i}\cos\varphi + \widehat{j}\,\mathrm{sen}\,\varphi\right) - \widehat{k}\,\mathrm{sen}\,\theta \\
\widehat{\varphi} &= -\widehat{i}\,\mathrm{sen}\,\varphi + \widehat{j}\cos\varphi \\
h_r &= 1 \qquad h_\theta = r \qquad h_\varphi = r\,\mathrm{sen}\,\theta \qquad (16)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{r} &= \widehat{i}\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)+\widehat{j}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2)+\widehat{k}\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_2|k_2)\\
\widehat{\chi}_1 &= (r|h_{\chi_1})\Big[-\widehat{i}k_1^2\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)\\
&\quad -\widehat{j}\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2)+\widehat{k}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{dn}\,(\chi_2|k_2)\Big]\\
\widehat{\chi}_2 &= (r|h_{\chi_1})\Big[\widehat{i}\,\mathrm{dn}\,(\chi_1|k_1)\,\mathrm{cn}\,(\chi_2|k_2)\,\mathrm{dn}\,(\chi_2|k_2)\\
&\quad -\widehat{j}\,\mathrm{cn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)\,\mathrm{dn}\,(\chi_2|k_2)-k_2^2\widehat{k}\,\mathrm{sn}\,(\chi_1|k_1)\,\mathrm{sn}\,(\chi_2|k_2)\,\mathrm{cn}\,(\chi_2|k_2)\Big]\\
h_{\chi_1} &= h_{\chi_2}=r\sqrt{1-k_1^2\,\mathrm{sn}^2\,(\chi_1|k_1)-k_2^2\,\mathrm{sn}^2\,(\chi_2|k_2)}
\end{aligned}
\tag{17}
$$

Nótese que los conjuntos de vectores unitarios $\left(\widehat{r},\widehat{\theta},\widehat{\varphi}\right)$ y $(\widehat{r},\widehat{\chi}_1,\widehat{\chi}_2)$ forman triadas ortonormales a mano derecha.

Los operadores de gradiente toman las formas

$$
\begin{aligned}
\nabla &= \widehat{r}\frac{\partial}{\partial r}+\frac{\widehat{\vartheta}}{r}\frac{\partial}{\partial\theta}+\frac{\widehat{\varphi}}{r\,\mathrm{sen}\,\theta}\frac{\partial}{\partial\varphi}\\
&= \widehat{r}\frac{\partial}{\partial r}+\frac{\widehat{\chi}_1}{h_{\chi_1}}\frac{\partial_1}{\partial\chi_1}+\frac{\widehat{\chi}_2}{h_{\chi_2}}\frac{\partial}{\partial\chi_2}
\end{aligned}
\tag{18}
$$

en coordenadas esféricas y esferoconales respectivamente.

La divergencia de un campo vectorial tiene la forma general

$$
\nabla\cdot\overrightarrow{\vee}=\frac{1}{h_1h_2h_3}\left[\frac{\partial}{\partial q_1}\left(h_2h_3\vee_1\right)+\frac{\partial}{\partial q_2}\left(h_1h_3\vee_2\right)+\frac{\partial}{\partial q_3}\left(h_1h_2\vee_3\right)\right] \tag{19}
$$

tomando en cuenta los elementos diferenciales de volumen $d\tau=h_1h_2h_3dq_1dq_2dq_3$, y los elementos de área $\overrightarrow{e}_1h_1dq_2dq_3+\overrightarrow{e}_2h_1h_3dq_2+\overrightarrow{e}_3h_1h_2dq_1dq_2$ necesarios para satisfacer el teorema de Gauss.

Correspondientemente, el operador de Laplace en las coordenadas de nuestro interés, toma las formas respectivas:

$$
\begin{aligned}
\nabla^2 &= \frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r}+\frac{1}{r^2}\left[\frac{1}{\mathrm{sen}\,\theta}\frac{\partial}{\partial\theta}\,\mathrm{sen}\,\theta\frac{\partial}{\partial\theta}+\frac{1}{\mathrm{sen}^2\,\theta}\frac{\partial^2}{\partial\varphi^2}\right]\\
&= \frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r}+\frac{1}{h_\chi^2}\left[\frac{\partial^2}{\partial\chi_1^2}+\frac{\partial^2}{\partial\chi_2^2}\right].
\end{aligned}
\tag{20}
$$

En mecánica cuántica las cantidades dinámicas se representan con operadores. En particular, la cantidad de movimiento lineal es proporcional al operador de gradiente,

$$\widehat{\overrightarrow{p}} = -i\hbar\nabla \tag{21}$$

y su cuadrado es proporcional al Laplaciano,

$$\widehat{\overrightarrow{p}} \cdot \widehat{\overrightarrow{p}} = -\hbar^2\nabla^2, \tag{22}$$

donde $\hbar$ es la constante de Planck reducida. El momento angular orbital está definido clásicamente por

$$\overrightarrow{L} = \overrightarrow{r} \times \overrightarrow{p} \tag{23}$$

y como vector es transversal tanto a $\overrightarrow{r}$ como a $\overrightarrow{p}$.

Para descomponer el vector $\overrightarrow{p}$ en sus partes radial y transversales multiplicamos vectorialmente el vector unitario radial por ambos miembros de la última ecuación

$$\widehat{r} \times \overrightarrow{L} = \widehat{r} \times (\overrightarrow{r} \times \overrightarrow{p}) = \overrightarrow{r}\widehat{r} \cdot \overrightarrow{p} - r\overrightarrow{p}, \tag{24}$$

con el resultado final

$$\overrightarrow{p} = \widehat{r}\left(\widehat{r} \cdot \overrightarrow{p}\right) - \frac{\widehat{r}\times\overrightarrow{L}}{r}. \tag{25}$$

La comparación de los operadores de las Ecs. (20) y (22) con los operadores asociados a la Ec. (25) permite reconocer que los operadores en (θ, φ) y (χ_1, χ_2) en la Ec. (20) corresponden al operador del cuadrado de momento angular

$$\widehat{\overrightarrow{L}} \cdot \widehat{\overrightarrow{L}} = -\hbar^2\left[\frac{1}{\text{sen}\theta}\frac{\partial}{\partial\theta}\,\text{sen}\;\theta\frac{\partial}{\partial\theta} + \frac{1}{\text{sen}^2\theta}\frac{\partial^2}{\partial\varphi^2}\right] = -\frac{\hbar^2 r^2}{h_x^2}\left[\frac{\partial^2}{\partial\chi_1^2} + \frac{\partial^2}{\partial\chi_2}\right]. \tag{26}$$

También es útil escribir las componentes cartesianas del operador de momento angular usando las ecs. (21) y (23):

$$\begin{aligned}
\widehat{L}_x &= -i\hbar\left(y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}\right) \\
\widehat{L}_y &= -i\hbar\left(z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}\right) \\
\widehat{L}_z &= -i\hbar\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right).
\end{aligned} \tag{27}$$

El lector puede usar las reglas de conmutación

$$[x_i, p_i] = i\hbar \tag{28}$$

para establecer las siguientes reglas de conmutación

$$\left[\widehat{L}_x, \widehat{L}_y\right] = i\hbar\widehat{L}_z \tag{29}$$

con cambios cíclicos de x, y, z.y adicionalmente,

$$\left[\widehat{L}^2, \widehat{L}_i\right] = 0 \tag{30}$$

para $i = x, y, z$; así como

$$\left[\widehat{L}^2, \widehat{p}^2\right] = 0, \left[\widehat{p}^2, \widehat{L}_z\right] = 0. \tag{31}$$

Las anulaciones de los tres últimos conmutadores indica que los operadores $\widehat{L}^2, \widehat{p}^2$ *y* $\widehat{L}_z$ tienen eigenfunciones comunes que son los armónicos esféricos familiares. Los conmutadores de las Ecs. (29) y similares indican que en general las eigenfunciones de $\widehat{L}_z$ no pueden ser eigenfunciones de $\widehat{L}_x$ y $\widehat{L}_y$.

El Hamiltoniano de las rotaciones de moléculas asimétricas, en el sistema fijo en el cuerpo y orientado a lo largo de sus ejes principales, queda expresado en términos de los cuadrados de las componentes del momento angular y los momentos de inercia respectivos I_1, I_2, I_3:

$$\widehat{H} = \left[\frac{\widehat{L}_x^2}{I_1} + \frac{\widehat{L}_y^2}{I_2} + \frac{\widehat{L}_z^2}{I_3}.\right] \tag{32}$$

Una parametrización alternativa introduce la traza promediada de la matriz de los inversos de los momentos de inercia,

$$Q = \frac{1}{3}\left[\frac{1}{I_1} + \frac{1}{I_2} + \frac{1}{I_3}\right], \tag{33}$$

la magnitud de la asimetría de la molécula P y tres parámetros adimensionales e_i de distribución de la asimetría, tales que

$$Pe_i = \frac{1}{I_i} - Q \tag{34}$$

$$e_i + e_2 + e_3 = 0 \tag{35}$$

$$e_1^2 + e_2^2 + e_3^2 = \frac{3}{2} \tag{36}$$

$$P^2 = \frac{2}{9}\left[\left(\frac{1}{I_1} - \frac{1}{I_2}\right)^2 + \left(\frac{1}{I_1} - \frac{1}{I_3}\right)^2 + \left(\frac{1}{I_2} - \frac{1}{I_3}\right)^2\right] \tag{37}$$

El Hamiltoniano de la Ec. (32) queda reescrito como la combinación

$$\widehat{H} = \frac{1}{2}Q\ \widehat{L}^2 + \frac{1}{2}P\left(e_1^2\widehat{L}_x^2 + e_2\widehat{L}_y^2 + e_3\widehat{L}_z^2\right) \tag{38}$$

la energía de un trompo esférico y la energía de la parte asimétrica con magnitud P y distribución

$$\widehat{H}^* = \frac{1}{2}\left[e_1\widehat{L}_x^2 + e_2\widehat{L}_y^2 + e_3\widehat{L}_z^2.\right] \tag{39}$$

Los operadores $\widehat{H}, \widehat{L}^2$ *y* $\widehat{H}^*$ conmutan entre sí

$$\left[\widehat{H}, \widehat{L}^2\right] = 0, \left[\widehat{L}^2, \widehat{H}^*\right] = 0, \left[\widehat{H}, \widehat{H}^*\right] = 0 \tag{40}$$

y por lo tanto comparten eigenfunciones comunes, que son los armónicos esferoconales.

Aquí incluimos también la forma explícita del Hamiltoniano de la Ec. (39), la cual puede construirse a partir de las Ecs. (27) para los componentes del momento angular y las Ecs. (9) de transformación entre coordenadas cartesianas y esferoconales:

$$\widehat{H}^* = -\frac{\hbar^2\, r^2}{2\, h_\chi^2}\left\{\left[e_1 - (e_1 - e_2)\,\mathrm{sn}^2\left(\chi_1|k_1\right)\right]\frac{\partial^2}{\partial\chi_1^2} + \left[e_3 - (e_2 - e_3)\,\mathrm{sn}^2\left(\chi_2|k_2\right)\right]\frac{\partial^2}{\partial\chi_2^2}\right\} \tag{41}$$

El lector puede apreciar las similaridades y diferencias con respecto a la Ec. (26).

3 Funciones Armónicas cartesianas, esféricas y esferoconales. La ecuación de Laplace en coordenadas cartesianas,

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\Phi\left(x,y,z\right)=0 \tag{42}$$

admite como soluciones las funciones constante 1; lineales x, y, z; cuadráticas xy, xz, yz, y cúbica xyz. Obviamente, todas tienen derivadas parciales de segundo orden que se anulan en la Ec. (8), y por lo tanto son funciones armónicas.

A continuación analizamos el efecto de los operadores de momento angular de las Ecs. (27) sobre las funciones sucesivas:

$$\widehat{L}_i 1 = 0 \tag{43}$$

para $i = x, y, z$. La función armónica 1 es isotrópica, eigenfunción de las tres componentes del momento angular con el eigenvalor común cero.

	$\widehat{L}_x/\left(-i\hbar\right)$	$L_y/\left(-i\hbar\right)$	$L_z/\left(-i\hbar\right)$
x	0	z	$-y$
y	$-z$	0	x
z	y	$-x$	0

La aplicación de los operadores sobre las funciones lineales da cero para la coordenada en la dirección de la componente, y la coordenada transversal a la componente del operador y a la coordenada sobre la que se aplica. En otras palabras, las coordenadas a lo largo de la componente del operador no se ve afectada bajo una rotación alrededor de la dirección de su eje común, mientras que las coordenadas transversales bajo la misma operación se transforman la una en la otra, como se aprecia en cada columna.

	$L_x/\left(-i\hbar\right)$	$L_y/\left(-i\hbar\right)$	$L_z/\left(-i\hbar\right)$
yz	y^2-z^2	$-xy$	xz
zy	yz	z^2-x^2	$-yz$
xy	$-zx$	zy	x^2-y^2
x^2-y^2	$2zy$	$2zx$	$-4xy$
$2z^2-x^2-y^2$	$6yz$	$-6zx$	0

La acción de los tres operadores sobre cada una de las tres funciones originales en los tres renglones superiores conduce a funciones armónicas diferencias de los cuadrados de las coordenadas involucradas para los resultados en la diagonal, y a las otras dos funciones cuadráticas originales para los resultados fuera de la diagonal. La función en el cuarto renglón es el resultado en el tercer renglón de la tercera columna, y bajo la acción de los operadores conduce a las tres funciones originales, respectivamente. El quinto renglón es también una función armónica formada como la diferencia del resultado del segundo renglón de la segunda columna menos el resultado del primer renglón primera columna. La acción de los dos primeros operadores sobre

esta función conduce a las dos primeras funciones originales, y aplicarle el tercer operador es anularlo como lo indica el quinto renglón de la tercera columna.

Los resultados correspondientes para las funciones armónicas cúbicas son las siguientes:

	$L_x/(-i\hbar)$	$L_y/(-i\hbar)$	$L_z/(-i\hbar)$
xyz	$x(y^2-z^2)$	$y(z^2-x^2)$	$z(x^2-y^2)$
$(x^2-y^2)z$	$y(x^2-y^2+2z^2)$	$x(2z^2-x^2+y^2)$	$-4xyz$
$x(3y^2-x^2)$	$-6xzy$	$-3(x^2-y^2)z$	$3y(3x^2-y^2)$
$y(3x^2-y^2)$	$-3(x^2-y^2)z$	$6xyz$	$3x(x^2-3y^2)$
$x(4z^2-x^2-y^2)$	$10xyz$	$(4z^2-11x^2-y^2)z$	$-y(4z^2-x^2-y^2)$
$y(4z^2-x^2-y^2)$	$-(4z^2-x^2-11y^2)z$	$-10xyz$	$x(4z^2-x^2-y^2)$
$(2z^2-3x^2-3y^2)z$	$3y(4z^2-x^2-y^2)$	$-3(4z^2-x^2-y^2)$	0

Partiendo de la función original xyz, el lector puede verificar directamente que los resultados de aplicarle los operadores de momento angular en las tres columnas del primer renglón satisfacen la ecuación de Laplace. La función armónica en el segundo renglón se toma del resultado del primer renglón y tercera columna. Los resultados de aplicarle los operadores respectivos en las columnas uno y dos exhiben la misma forma con el intercambio de los papeles de x y y, y en la columna tres corresponde a la función original. Las funciones armónicas de los renglones tres y cuatro se obtienen al eliminar la dependencia en z al formar la combinación lineal de dos veces el resultado del primer renglón y primera columna más el resultado del segundo renglón y segunda columna; y también la combinación lineal del resultado del segundo renglón y primera columna menos dos veces el resultado del primer renglón y segunda columna, respectivamente. Los resultados de aplicar los operadores de las dos primeras columnas son las funciones armónicas de los dos primeros renglones, y en la tercera columna se intercambian las funciones armónicas tercera y cuarta. Las funciones armónicas de los renglones quinto y sexto resultan de combinar las mismas funciones del par anterior con la condición de que x^2 y y^2 tengan los mismos coeficientes. Al aplicarles el operador de la tercera columna, ambas funciones se intercambian como lo hicieron los dos pares anteriores. En los renglones quinto y sexto en las columnas uno y dos, vuelve a aparecer la función de partida xyz y dos combinaciones con los papeles de x^2y y^2 intercambiados. La séptima función armónica se forma restando las dos últimas y eliminando el factor común de 4. Al aplicarle el operador de la primera columna se reproduce la función número seis, el operador de la segunda columna lo convierte en la función número cinco, y el de la tercera columna lo anula como en los tres casos anteriores.

Las funciones armónicas x, y, z tienen sus representaciones esféricas y esferoconales por medio de las Ecs. (9) de transformación entre los respectivos sistemas de coordenadas. Para las funciones armónicas cuadráticas y cúbicas se identifican sus formas separables en coordenadas esféricas y esferoconales

$$xz = r^2 \operatorname{sen}\theta\cos\theta\cos\varphi = r^2 \operatorname{dn}(\chi_1|k_1)\operatorname{sn}(\chi_1|k_1)\operatorname{sn}(\chi_2|k_2)\operatorname{dn}(\chi_2|k_2)$$

$$yz = r^2 \operatorname{sen}\theta\cos\theta\operatorname{sen}\varphi = r^2 \operatorname{cn}(\chi_1|k_1)\operatorname{sn}(\chi_1|k_1)\operatorname{cn}(\chi_2|k_2)\operatorname{dn}(\chi_2|k_2)$$

$$xy = \tfrac{1}{2}r^2 \operatorname{sen}^2\theta\operatorname{sen}2\varphi = r^2 \operatorname{dn}(\chi_1|k_1)\operatorname{cn}(\chi_1|k_1)\operatorname{sn}(\chi_2|k_2)\operatorname{cn}(\chi_2|k_2)$$

$$x^2-y^2=r^2\,\mathrm{sen}^2\,\theta\cos 2\varphi$$
$$2z^2-x^2-y^2=3z^2-r^2=r^2\left(3\cos^2\theta-1\right) \tag{44}$$

$$xyz=\tfrac{1}{2}r^3\,\mathrm{sen}^2\,\theta\cos\theta\,\mathrm{sen}\,2\varphi=r^3\mathrm{dn}\left(\chi_1|k_1\right)\mathrm{cn}\left(\chi_1|k_1\right)\mathrm{sn}\left(\chi_1|k_1\right)$$
$$sn\left(\chi_2|k_2\right)\mathrm{cn}\left(\chi_2|k_2\right)\mathrm{cn}\left(\chi_2|k_2\right)\mathrm{dn}\left(\chi_2|k_2\right)$$
$$\left(x^2-y^2\right)z=r^3\,\mathrm{sen}^2\,\theta\cos\theta\cos 2\varphi$$
$$x\left(3y^2-x^2\right)=r^3\,\mathrm{sen}^3\,\theta\cos 3\varphi$$
$$y\left(3x^2-y^2\right)=r^3\,\mathrm{sen}^3\,\theta\,\mathrm{sen}\,3\varphi$$
$$x\left(4z^2-x^2-y^2\right)=r^3\left(5\cos^2\theta-1\right)\mathrm{sen}\,\theta\cos\varphi$$
$$y\left(4z^2-x^2-y^2\right)=r^3\left(5\cos^2\vartheta-1\right)\mathrm{sen}\,\theta\,\mathrm{sen}\,\varphi$$
$$\left(2z^2-3x^2-3y^2\right)z=r^3(5\cos^3\theta-3\cos\theta) \tag{45}$$

El lector puede identificar que estas funciones armónicas tienen paridades bien definidas ($x\to -x,\ \ y\to -y, z\to -z$). En la representación esférica son eigenfunciones de $\widehat{L}_z^2$. La fórmula de Euler $e^{im\varphi}=\cos m\varphi+i\ \mathrm{sen}\,m\varphi$ permite construir las combinaciones de los pares respectivos que son eigenfunciones de $\widehat{L}_z$,pero ya no tienen paridad definida.

Las tres primeras funciones armónicas cuadráticas y la primera función armónica cúbica son separables en coordenadas esferoconales. A continuación se ilustra la construcción de los armónicos esferoconales con paridad $(+,+,+)$ que son las contrapartes del último par de funciones armónicas esféricas cuadráticas.

La construcción requiere la solución simultánea de las ecuaciones de eigenvalores de los operadore de las Ecs., (20) y (41):

$$-\tfrac{\hbar^2}{1-k_1^2\mathrm{sn}^2(\chi_1|k_1)-k_2\mathrm{sn}^2(\chi_2|k_2)}\left[\tfrac{\partial^2}{\partial x_1^2}+\tfrac{\partial^2}{\partial x_2^2}\right]\Psi\left(\chi_1,\chi_2\right)=\hbar^2\ell\left(\ell+1\right)\Psi\left(\chi_1,\chi_2\right) \tag{46}$$

$$-\tfrac{\hbar^2}{2\left[1-k_1^2\mathrm{sn}^2(\chi_1|k_1)-k_2^2\mathrm{sn}^2(\chi_2|k_2)\right]}\Big\{e_1-\left(e_1-e_2\right)\mathrm{sn}^2\left(\chi_1|k_1\right)\tfrac{\partial}{\partial\chi_1^2}+$$

$$+\left[e_3-\left(e_2-e_3\right)\mathrm{sn}^2\left(\chi_2|k_2\right)\right]\tfrac{\partial^2}{\partial x_2^2}\Big\}\Psi\left(\chi_2|k_2\right)=E^*\Psi\left(\chi_1,\chi_2\right) \tag{47}$$

Efectivamente, ambas admiten la solución factorizable

$$\Psi\left(\chi_1,\chi_2\right)=X_1\left(\chi_1\right)X_2\left(\chi_2\right)$$

siempre y cuando los parámetros geométricos k_1 y k_2 se conecten con los parámetros de distribución de asimetría e_1, e_2 y e_3 como

$$k_1^2=\tfrac{e_2-e_3}{e_1-e_3}, k_2^2=\tfrac{e_1-e_2}{e_1-e_3}, \tag{48}$$

y con lo cual las dos ecuaciones diferenciales parciales acopladas conducen a las ecuaciones diferenciales ordinarias

$$\frac{d^2X_1(\chi_1)}{d\chi_1^2} - \left[\ell\left(\ell+1\right)k_1^2\mathrm{sn}^2\left(\chi_1|k_1\right) + \frac{\ell(\ell+1)e_3}{e_1-e_3} - \frac{2E^*}{\hbar^2(e_1-e_3)}x\right]X_1\left(\chi_1\right) = 0 \tag{49}$$

$$\frac{d^2X_2(\chi_2)}{d\chi_2^2} - \left[\ell\left(\ell+1\right)k_2^2\mathrm{sn}^2\left(\chi_2|k_2\right) + \frac{\ell(\ell+1)e_1}{e_1-e_3} + \frac{2E^*}{\hbar^2(e_1-e_3)}x\right]X_2\left(\chi_2\right) = 0. \tag{50}$$

Ambas corresponden a la ecuacion diferencial de Lamé en la forma estandar

$$\frac{d^2X_i(\chi_i)}{d\chi_i^2} - \left[\ell\left(\ell+1\right)k_i^2\mathrm{sn}^2\left(\chi_i|k_i\right) + h_i\right]X_i\left(\chi_i\right) = 0 \tag{51}$$

Los eigenvalores respectivos se identifican de inmediato,

$$h_1 = \frac{\ell(\ell+1)e_3}{e_1-e_3} - \frac{2E^*}{\hbar^2\left(e_1-e_3\right)}$$

$$h_2 = -\frac{\ell(\ell+1)e_1}{e_1-e_3} + \frac{2E^*}{\hbar^2\left(e_1-e_3\right)}, \tag{52}$$

y determinan los eigenvalores del cuadrado del momento angular y de la energía de rotación de moléculas asimétricas, respectivamente:

$$\ell\left(\ell+1\right) = -(h_1+h_2)$$

$$2E^*/\hbar^2 = -\left(e_1h_1 + e_3h_2\right) \tag{53}$$

La solución de la ecuación diferencialde Lamé (52) se ilustra a continuación para el caso más sencillo no trivial en que $\ell = 2$, y estados de paridad positiva en que la eigenfunción toma la forma:

$$X_i(\chi_i) = a_0 + a_1\mathrm{sn}^2\left(\chi_1|k_1\right). \tag{54}$$

Sus derivadas de primero y segundo orden se obtienen usando sucesivamente las Ecs. (12) , con los resultados,

$$\frac{dX_i}{d\chi_i} = a_1 2\mathrm{sn}\left(\chi_i|k_i\right)\mathrm{cn}\left(\chi_i|k_i\right)\mathrm{dn}\left(\chi_i|k_i\right)$$

$$\begin{aligned}\frac{d^2\chi_i}{d\chi_i^2} &= 2a_1\left[\mathrm{cn}^2\left(\chi_i|k_i\right)\mathrm{dn}^2\left(\chi_i|k_i\right) - \mathrm{sn}^2\left(\chi_i|k_i\right)\mathrm{dn}^2\left(\chi_i|k_i\right) - k_i^2\mathrm{sn}^2\left(\chi_i|k_i\right)\mathrm{cn}^2\left(\chi_i|k_i\right)\right] \\ &= 2a_1\left[1 - 2\left(1+k_i^2\right)\mathrm{sn}^2\left(\chi_i|k_i\right) + 3k_i^2\mathrm{sn}^4\left(\chi_i|k_i\right)\right],\end{aligned} \tag{55}$$

donde en el último renglón se usan las relaciones de las Ecs.(11) entre las diferentes funciones elípticas.

Al sustituir la función y su derivada segunda en la ecuación diferencial (52), se obtiene

$$2a_1\left[1 - 2\left(1+k_i^2\right)\mathrm{sn}^2\left(\chi_i|k_i\right) + 3k_i^2\mathrm{sn}^4\left(\chi_i|k_i\right)\right]$$

$$-\left[6k_i^2\mathrm{sn}^2\left(\chi_i|k_i\right)+h_i\right]\left[a_o+a_1\mathrm{sn}^2\left(\chi_i|k_i\right)\right]=0$$

Tomando en cuenta la independencia lineal de las potencias sucesivas de la función elíptica notamos la cancelación automática del coeficiente de la más alta, y las relaciones de recurrencia de los coeficientes de las potencias más bajas:

$$2a_1-h_ia_0=0$$

$$-4a_1\left(1+k_i^2\right)-h_ia_1-6k_i^2a_0=0.$$

Las razones de los coeficientes conducen a la ecuación que determina el eigenvalor h_i,

$$\tfrac{a_1}{a_0}=\tfrac{1}{2}h_i=-\tfrac{6k_i^2}{4\left(1+k_i^2\right)+hi}$$

Se trata de una ecuación de segundo grado,

$$h_i^2+4\left(1+k_i^2\right)h_i+12k_i^2=0$$

que admite dos soluciones

$$h_i^{\pm}=-2\left(1+k_i^2\right)\pm\sqrt{4\left(1+k_i^2\right)^2-12k_i^2}.$$

Los valores explícitos para $i=1,2$ son

$$h_i^{\mp}=-2\left(1+k_i^2\right)\pm\sqrt{4+8k_1^2+4k_i^4-12k_i^2}$$

$$=-2\left(1+k_i^2\right)\pm\sqrt{4-4k_i^2+\left(1-k_i^2\right)}$$

$$=-2\left(1+k_i^2\right)\pm2\sqrt{1-k_i^2k_2^2}$$

$$h_i^{\mp}=-2\left(1+k_2^2\right)\mp2\sqrt{1-k_i^2k_2^2} \tag{56}$$

compatibles con la condición de que sus sumas son

$$h_i^{\pm}+h_i^{\mp}=-2\left(1+k_i^2\right)-2\left(1+k_2^2\right)=-6=-2\times3$$

Los armónicos esferoconales cuadrupolares ($\ell=2$) respectivos son

$$X\left(x_1,h_1^+\right)X\left(\chi_2,h_2^-\right)=a_0^+\left(1+\tfrac{1}{2}h_1^+\mathrm{sn}^2\left(\chi_i|k_i\right)\right)a_0^-\left(1+\tfrac{1}{2}h_2^-\mathrm{sn}^2\left(\chi_2|k_2\right)\right) \tag{57}$$

$$X\left(x_1,h_1^-\right)X\left(\chi_2,h_2^+\right)=a_0^-\left(1+\tfrac{1}{2}h_1^-\mathrm{sn}^2\left(\chi_i|k_i\right)\right)a_0^+\left(1+h_2^+\mathrm{sn}^2\left(\chi_2|k_2\right)\right) \tag{58}$$

Ambos se pueden expresar en la forma

$$\phi\left(x,y,x\right)=Ax^2+By^2+Cz^2$$

$$= A\left(1-k_1^2\mathrm{sn}^2\left(\chi_1|k_1\right)\right)\mathrm{sn}^2\left(\chi_2|k_2\right)+B\left(1-\mathrm{sn}^2\left(\chi_1|k_1\right)\right)\left(1-\mathrm{sn}^2\left(\chi_2|k_2\right)\right)+$$
$$C\mathrm{sn}^2\left(\chi_1|k_1\right)\left(1-k_2^2\mathrm{sn}^2\left(\chi_2|k_2\right)\right) \tag{59}$$

con la condición de ser armónica, lo que implica que

$$A+B+C=0$$

además de la igualdad de los coeficientes de las potencias sucesivas de las potencias de las funciones elípticas.

$$\begin{aligned} B &= a_0^+a_0^- \\ -B+C &= a_0^+a_0^-\frac{1}{2}h_1^+ \\ -B+A &= a_0^+a_0^-\frac{1}{2}h_2^- \\ B-k_1^2A-k_2^2C &= a_0^+a_0^-\frac{1}{4}h_1^\pm h_2^\mp. \end{aligned}$$

El lector puede verificar la compatibilidad de las condiciones mencionadas y reconocer las proporciones

$$A\colon B\colon C\colon :1+\frac{1}{2}h_2^\mp\colon 1\colon 1+\frac{1}{2}h_1^\pm.$$

Para el caso de las soluciones cúbicas, la función armónica original es separable tanto en coordenadas cartesianas y esferoconales:

$$xyz=X^{dcs}\left(\chi_1,h_1\right)X^{\mathrm{scd}}\left(\chi_2,h_2\right) \tag{60}$$

donde s,c,d son las inciales de las funciones elípticas que representan las singularidades factorizables de la ecuación diferencial de Lamé, y los eigenvalores son $\mathrm{h}_i^{dcs}=-4\left(1+k_1^2\right),i=1,2$. Existen además tres pares de soluciones de las formas

$$\begin{aligned} X^d\left(\chi_1,h_1^d\right)X^s\left(\chi_2,h_2^d\right) &= \mathrm{dn}\left(\chi_1|k_1\right)\mathrm{sn}\left(\chi_2|k_2\right)\left[a_0^d+a_1^d\,\mathrm{sn}^2\left(\chi_1|k_1\right)\right]\left[a_0^s+a_1^s\,\mathrm{sn}^2\left(\chi_2|k_2\right)\right] \\ X^c\left(\chi_1,h_1^c\right)X^c\left(\chi_2,h_2^c\right) &= \mathrm{cn}\left(\chi_1|k_1\right)\mathrm{cn}\left(\chi_2|k_2\right)\left[a_0^c+a_1^c\,\mathrm{sn}^2\left(\chi_1|k_1\right)\right]\left[a_0^c+a_1^c\,\mathrm{sn}^2\left(\chi_2|k_2\right)\right] \\ X^s\left(\chi_1,h_1^s\right)X^d\left(\chi_2,h_2^d\right) &= \mathrm{sn}\left(\chi_1|k_1\right)\mathrm{dn}\left(\chi_2k_2\right)\left[a_0^s+a_1^s\,\mathrm{sn}^2\left(\chi_1|k_1\right)\right]\left[a_0^d+a_1^d\,\mathrm{sn}^2\left(\chi_2|k_2\right)\right] \end{aligned} \tag{61}$$

análogas al último par de las funciones armónicas cuadráticas.

En los tres pares sucesivos bajo consideración se reconocen los factores asociados a x,y,z, Ec. (9), que definen las paridades negativas de las funciones armónicas respectivas.

4 Discusión

En las secciones 2 y 3 se ha presentado la construcción sistemática de los operadores Laplaciano, de momento angular y el Hamiltoniano de rotaciones de moléculas asimétricas, e ilustrado la identificación de sus soluciones comunes, funciones armónicas y eigenfunciones de momento angular, en coordenadas cartesianas, esféricas y esferoconales. Se aprovecha la familiaridad con las dos primeras para guiar al lector a conocer las terceras. Los lectores están invitados a ampliar su conocimiento sobre las mismas por medio de la referencia [1] y de los trabajos ahí citados.

Algunas aplicaciones de los armónicos esferoconales en diferentes áreas están siendo investigadas por el autor y sus colaboradores. Resultados preliminares sobre algunas de ellas se reportarán en el XLIV Congreso Nacional de Física:

1) Campos Electrostáticos y Magnetostáticos Multipolares Esferoconales Completos, Internos y Externos, y sus Fuentes sobre una Esfera.

2) Campos Magnéticos con Gradiente Constante y sus Embobinados en Base de Armónicos Esferoconales.

3) Interacciones Cuadrupolares en Bases de Armónicos Esferoconales.

4) Momentos de Inercia de Moléculas Asimétricas obtenidos del Análisis Armónico Esferoconal de sus Espectros de Rotación Experimentales.

5 Referencias

[1]. R. Méndez-Fragoso y E. Ley-Koo, Advances in Quantum Chemistry, Vol. 62, Cap. 4, 137-213(2011).

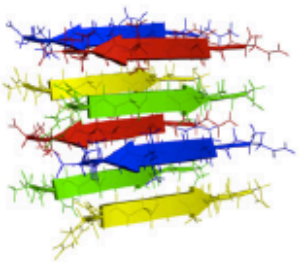
[2]. M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dever, New York, Capítulo 16 (1965).

Viaje por el paisaje conformacional de una proteína con doble personalidad

Nina Pastor
Facultad de Ciencias, UAEM
nina@uaem,mx

Los avances en medicina y tecnología están haciendo más longeva a la especie humana. La población humana mundial se está haciendo vieja, y esto hace interesante y urgente entender todo un grupo de enfermedades que aquejan típicamente a personas mayores de 60 años. Además del cáncer y los problemas cardiovasculares y metabólicos, existe un grupo de enfermedades causadas por el plegamiento anómalo de proteínas. La más famosa es la enfermedad de Alzheimer, por su devastador efecto sobre algo que consideramos muy humano: nuestras capacidades cognitivas.

Esta historia comienza con un problema médico que aqueja a 3000 personas por año en los Estados Unidos de América, llamado amiloidosis de cadena ligera, o AL (Merlini G and Bellotti V (2003) New Eng. J. Med. 349:583). Los pacientes se presentan con problemas en el corazón, hígado, riñones y/o tubo digestivo; algunos tienen dolor en las articulaciones. Al analizar la orina y el suero sanguíneo de estos pacientes, se encuentra una proteína en exceso, y ésta misma se encuentra depositada en los órganos que no funcionan bien. El material depositado está organizado en forma ordenada: se une a una molécula llamada rojo Congo, que genera birrefringencia verde cuando se ve con luz polarizada. Al examinar los depósitos con microscopía electrónica, se observan fibras delgadas y largas, sin ramificaciones. Estas son las fibras amiloides. Aumentando la resolución con crioelectromicroscopía, se puede ver que las fibras están retorcidas, y que el grado de enrollamiento depende de las condiciones de preparación de las fibras. A nivel atómico, todas las fibras dan un patrón de difracción de rayos X con dos señales típicas, una que indica repetición cada 4.7Å y otra a una distancia ~10Å. Este patrón se conoce como β-cruzada, y se obtiene con un sándwich de hebras β paralelas o antiparalelas entre sí; estos sándwiches luego pueden apilarse. A esta estructura también se le conoce como zipper estérico, porque las cadenas laterales de los aminoácidos se engarzan como en una cremallera en el interior del sándwich.



¿Por qué se forman las fibras amiloides? La idea biofísica más actual considera que adyacente al embudo de plegamiento de las proteínas coexiste un embudo de agregación. En el primero privan las interacciones intramoleculares, es decir, de la proteína consigo misma, mientras que en el segundo lo que importa son las interacciones entre proteínas. El embudo de plegamiento suele tener un mínimo de energía libre notablemente más profundo que todos los demás, correspondiente al estado nativo. Este estado no hace

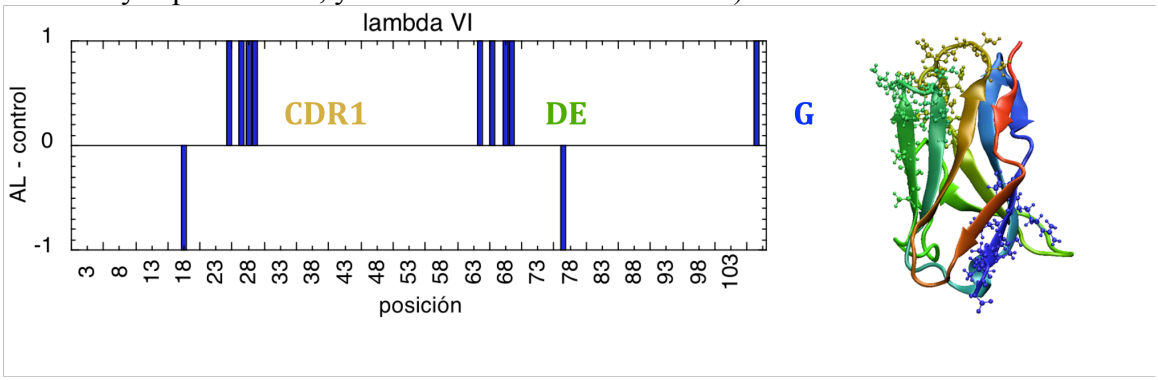
fibras, ya que ha satisfecho de manera óptima las interacciones intramoleculares y aquellas con el ambiente en el que está. Sin embargo, existen otros mínimos locales en el embudo, que corresponden a intermediarios de plegamiento. Estas especies parcialmente plegadas sí pueden formar fibras, evento que se vuelve más probable cuando la concentración de proteína aumenta. El embudo de agregación se supone mucho más rugoso que el de plegamiento, y esto explica el polimorfismo observado para las fibras. Para no dejar en el lector la idea de que las fibras amiloides son siempre patológicas, recientemente se han descrito muchos casos de amiloides funcionales. Al parecer, estas estructuras representan el universo de conformaciones que adquirían las proteínas más primitivas. Son una reliquia evolutiva (Eichner T and Radford SE (2011) Mol. Cell 43:8).

La proteína que forma las fibras en los pacientes con AL es un fragmento de anticuerpo. En condiciones normales, los anticuerpos son moléculas formadas por dos cadenas pesadas y dos cadenas ligeras, formando una estructura en forma de Y. En las puntas superiores de la Y se encuentran los dominios variables, que son los encargados de reconocer todo aquello que no somos nosotros (parásitos, por ejemplo). Como los anticuerpos tienen que ser capaces de unirse a una variedad muy grande de moléculas ajenas a nosotros, se construyen mediante un mecanismo combinatorio. Los humanos tenemos dos tipos de cadena ligera (κ y λ), compuestas de un dominio variable y uno constante. Los dominios variables se construyen eligiendo al azar un gene V y uno J. Tenemos del orden de 70 genes V, y una decena de genes J. Una vez que se ha unido un gene V con un gene J, tenemos lo que se conoce como un dominio variable rearreglado, pero inmaduro. Aquellas células del sistema inmune que eligieron una combinación útil para combatir al parásito del momento, reciben una señal de proliferación. Con cada ronda de replicación, incluyen mutaciones en este dominio variable. Las mutaciones tienen el propósito de mejorar la afinidad del anticuerpo por el agente externo, y son seleccionadas con este criterio (los que generan mutaciones que no mejoran la afinidad se replican menos, y acaban desapareciendo de la población). Lo que sucede en los pacientes con AL es que en lugar de producir un anticuerpo completo, las células que secretan anticuerpos exportan sólo las cadenas ligeras. El análisis de la secuencia de aminoácidos de muchos pacientes revela que las proteínas depositadas provienen de anticuerpos maduros. Se propone que el efecto de las mutaciones requeridas para madurar al anticuerpo desestabilizan al dominio variable, haciendo que éste se vuelva propenso a agregarse en forma de fibras amiloides. No todos los genes V y J aparecen en los casos clínicos. Llaman la atención dos clases, la 6a y la 3r, las cuales causan ~40% de los casos reportados, a pesar de ser sólo el 2% de los anticuerpos circulantes en personas sanas (Perfetti V *et al* (2002) Blood 100:948). Esto lleva a pensar que hay algo particularmente inestable en estos genes. Nosotros trabajamos con la clase 6a; el equipo de trabajo está formado por el laboratorio del Dr. Alejandro Fernández en la Facultad de Medicina de la UNAM, el Dr. César Millán en la UAM-Iztapalapa, y por mi laboratorio.

¿Cómo es la estructura de los dominios variables de inmunoglobulina? Resulta que este dominio es muy antiguo, ya que se usa para mediar interacciones entre células desde hace muchos miles de millones de años. Está compuesto por ~110 aminoácidos organizados mayormente como hebras β, dispuestas en dos planos paralelos entre sí. A esta arquitectura se le conoce como sándwich β. Recordemos que las fibras amiloides también

están hechas de hebras β. Para evitar la agregación lateral de estos dominios variables, las estructuras presentan lo que Richardson y Richardson (Richardson JS and Richardson DC (2002) Proc. Natl. Acad. Sci. USA 99:2754) han llamado mecanismos moleculares anti-agregación. Tienen rizos que ocluyen el borde del sándwich, y/o introducen curvatura en los bordes, para sellarlos y evitar que presenten donadores y aceptores de puente de hidrógeno para interacciones intermoleculares. El problema bioquímico al que nos enfrentamos es que no hay grandes diferencias en las estructuras de los dominios variables de pacientes con AL, y los de personas sanas. Además, recordemos que cada paciente sintetiza una proteína diferente. ¿Cómo podemos averiguar el efecto de estas variantes en secuencia sobre el comportamiento del dominio?

Un enfoque bioinformático que seguimos fue preguntar qué tan fácil es acomodar regiones pequeñas, de seis aminoácidos de longitud, provenientes de proteínas de pacientes con AL, en la estructura básica, central de la fibra amiloide, el zipper estérico. Para esto, usamos el servidor del programa ZipperDB (Goldschmidt L *et al.* (2010) Proc. Natl. Acad. Sci. USA 107:3487), el cual está calibrado para reportar como probable formador de fibra a todo segmento que tenga una energía potencial inferior a -23 kcal/mol. Aplicamos este programa a una colección de más de cien proteínas derivadas de pacientes con AL, y a controles sin la enfermedad. Buscamos las regiones del dominio que consistentemente daban una alta propensión a formar fibras en los pacientes y no en los controles. Encontramos tres regiones para las cadenas del tipo 6a: uno de los rizos encargado de reconocer antígenos (región CDR1), y dos de los bordes del dominio (las hebras D y E por un lado, y la hebra G en el otro extremo).



En cada paciente pueden involucrarse regiones diferentes, e incluso en un mismo paciente pudieran reclutarse distintas regiones del dominio para hacer fibras en distintos órganos. Es importante recalcar que en el estado nativo estas zonas pro-fibra no son accesibles al solvente; están amarradas al resto del dominio con los mecanismos de seguridad ya descritos. De este análisis se desprende la siguiente hipótesis obligada: los pacientes con AL debilitan estos mecanismos de seguridad, haciendo más accesibles las regiones pro-fibra del dominio.

Para ejemplificar el trabajo que realizamos en el laboratorio, ahora nos vamos a centrar en dos cadenas tipo 6a que sólo difieren en un aminoácido. Las llamaremos **R24** y **G24**. La primera variante ocurre en el 75% de la población humana, y la segunda, alrededor de 22%, por lo que son muy buenos modelos de cadenas rearregladas inmaduras humanas (del Pozo Yauner L *et al.* (2008) Proteins 72:684). Para estas dos proteínas contamos con

información estructural, ya que han sido cristalizadas (Hernández Santoyo A *et al* (2010) J. Mol. Biol. 396:280; Enrique Rudiño, comunicación personal). También tenemos medidas de su estabilidad termodinámica, y del tiempo que tardan en empezar a formar fibras. **G24** hace fibras 7 veces más rápido que **R24** (del Pozo Yauner L *et al.* (2008) Proteins 72:684). Mientras menos estable es la proteína, menos tiempo tarda en hacer fibras, y por lo tanto, esto sugiere que los oligómeros, núcleos y semillas de las fibras contienen estructuras parcial o totalmente desnaturalizadas. Nuestra misión es encontrar estos precursores parcialmente desnaturalizados; como viven poco tiempo y se dan a concentraciones bajísimas, son muy difíciles de caracterizar experimentalmente. La herramienta que usamos son las simulaciones por dinámica molecular.

Hacemos dos tipos de simulaciones. En el primer caso, nos interesa estudiar la topografía del mínimo de energía libre en el que reside el estado nativo, ya que se ha descrito que basta con desnaturalizar localmente a la proteína, generando un estado "excitado" llamado N^*, para poder saltar al embudo de agregación (Chiti F and Dobson CM (2009) Nature Chem. Biol. 5:15). Estas son simulaciones al equilibrio. En el segundo caso, queremos sacar a la proteína del estado nativo, y explorar su paisaje conformacional; éstas son simulaciones fuera del equilibrio. Ambos tipos de simulaciones las hacemos con el programa NAMD (Phillips JC *et al.* (2005) J. Comput. Chem. 26:1781) y el potencial clásico CHARMM27 (MacKerell Jr AD *et al.* (1998) The Encyclopedia of Computational Chemistry 1:271), en el ensamble NPT, con pasos de integración de 2 fs. La celda computacional se construye con el servidor CHARMM-GUI, usando como punto de partida las coordenadas cristalográficas de la proteína. Agregamos ~19000 moléculas de agua, y suficiente KCl para neutralizar la carga de la proteína y lograr una concentración 0.1M de sal. Esta celda es el punto de partida para todas las simulaciones, las cuales difieren en la temperatura de simulación y en la semilla de número aleatorio usada para asignar velocidades atómicas. El muestreo al equilibrio lo realizamos con cinco simulaciones de 50 ns cada una a 298K y otras cinco de 100 ns cada una a 398K. Se ha visto que cinco simulaciones dan un mejor muestreo que una sola cinco veces más larga. Las rutas de desnaturalización las estudiamos con diez simulaciones de 100 ns cada una a 498K, cantidad suficiente para estudios cualitativos.

El producto inmediato de todas las simulaciones es una colección de coordenadas atómicas ordenadas en el tiempo, a intervalos de 100 fs. A partir de estas coordenadas calculamos propiedades que nos permiten conectar el mundo de las simulaciones con las mediciones experimentales, y que nos ayudan a describir los movimientos de la proteína:

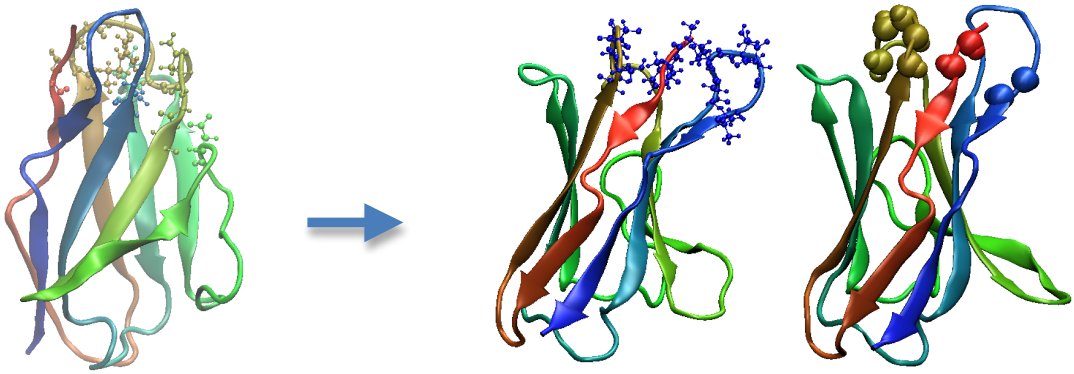
simulaciones	experimentos
radio de giro	radio de Stokes (filtración)
desplazamiento cuadrático medio	familias estructurales (cristalografía, resonancia)
fluctuaciones	factores B cristalográficos
distancias	FRET, NOE, PRE, fluorescencia
parámetros de orden	resonancia magnética nuclear
áreas superficiales	solvatación, fluorescencia
ángulos diedros	dicroísmo circular
energía interna	
contactos interatómicos	

La topografía de la cuenca nativa: cazando a N^*

¿De qué depende la rugosidad del mínimo de energía libre en el que reside el estado nativo? Cada mínimo local corresponde a un punto de equilibrio. Si todas las interacciones entre los átomos de la proteína son consistentes entre sí, entonces estamos en lo que Nobuhiro Gō llamó un estado de frustración mínima; esto corresponde a cuencas lisas, con un solo mínimo. Por el contrario, si hay grupos de aminoácidos cuyos contactos óptimos son incompatibles con los de otra zona, entonces tendremos a la proteína oscilando entre estas dos maneras de organizarse, cada una correspondiente a un mínimo local dentro de la cuenca nativa. Para saber cuál de estos escenarios se aplica mejor a **R24** y a **G24**, obtuvimos 2500 estructuras generadas a 298K para cada proteína (espaciadas 100 ps entre sí), y calculamos la raíz del desplazamiento cuadrático medio para cada par de átomos equivalentes (RMSD en inglés) para cada pareja de proteínas. Esto nos dice qué tan diferentes son entre sí las 2500 estructuras de cada simulación. Las agrupamos usando un radio de corte de 2Å. Si se trata de una estructura mínimamente frustrada, esperamos poder describir toda la muestra con un solo grupo; si hay frustración, esperamos encontrar cuantos grupos se requieran para describir los mínimos locales visitados durante las simulaciones. En estas condiciones encontramos que **G24** aparentemente está más frustrada que **R24**, ya que la primera visita tres mínimos locales y la segunda se puede describir con una sola estructura:

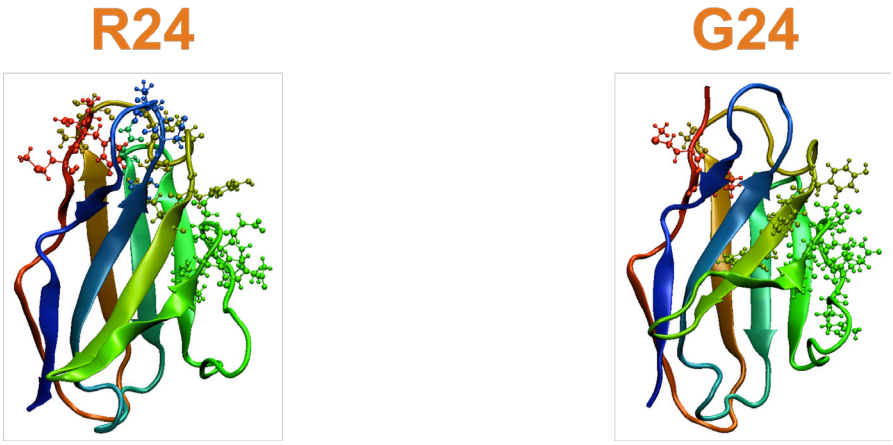
Todas estas estructuras contienen ~ 90% de los contactos interatómicos que definen al estado nativo, por lo que ninguna corresponde a N^*.

¿Por qué está incómoda **G24**? El cambio de arginina por glicina supone una pérdida de interacciones, las cuales podrían haber dado al traste con el delicado balance que estabiliza a la proteína. De las simulaciones calculamos la colección de interacciones que definen al estado nativo: distancias entre carbonos $\leq$ 6Å, y puentes de hidrógeno. Comparando la lista de contactos nativos de **R24** con la de **G24**, encontramos que la mutación perturba una zona amplia de la molécula, generando movimiento extra que se puede caracterizar con fluctuaciones de carbonos alfa y parámetros de orden del enlace NH:



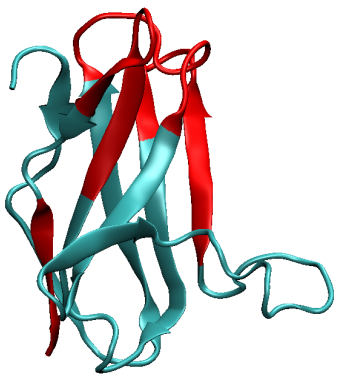
contactos perdidos → aumento en movimiento de carbonos alfa y enlaces NH

Para encontrar a N^* hay que aumentar la temperatura a 398K. Cada proteína pierde un grupo diferente de contactos, pero tienen algo en común: se pierden aquellos entre el asa C'-C" y el resto de la proteína. Este es uno de los mecanismos de protección anti-agregación, y por lo tanto, sugiere que éstos podrían ser intermediarios que conectan al embudo de plegamiento con el de agregación.

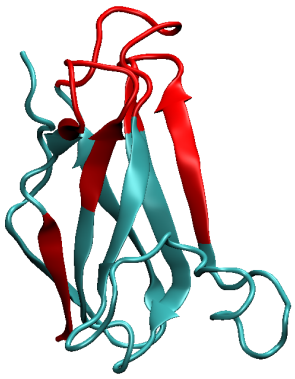


contactos nativos perdidos a 398K con respecto a 298K

Si agrupamos las estructuras visitadas durante las simulaciones a 398K para cada proteína, vemos que **G24** visita más mínimos locales que **R24**, y además, los puebla más frecuentemente. El 22% del tiempo **G24** expone la hebra D, la cual recordemos es una de las zonas pro-fibra en esta familia de cadenas ligeras. Por su parte, **R24** expone esta hebra también, pero con menor eficiencia (un 7% del tiempo). La siguiente figura muestra estructuras representativas de los grupos más poblados para **R24** y **G24**, con las zonas pro-fibra indicadas en rojo. Claramente se observa que se ha desplazado el asa C'-C", exponiendo a la hebra D. En el caso de **G24**, también se observa que el CDR1 está extendido hacia arriba, aumentando su exposición al solvente.



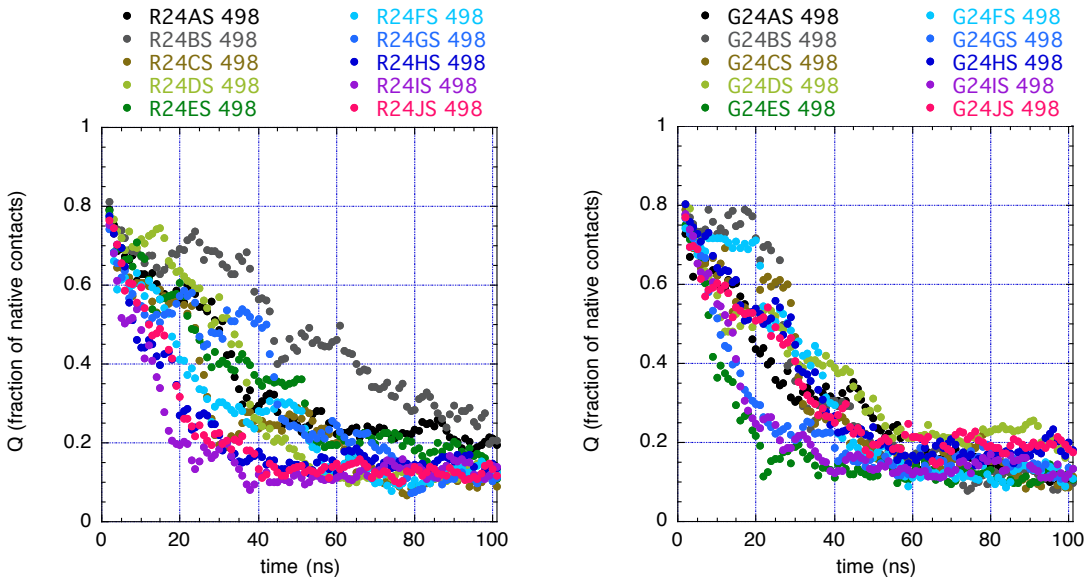
R24



G24

La salida de la cuenca nativa: múltiples rutas de desnaturalización

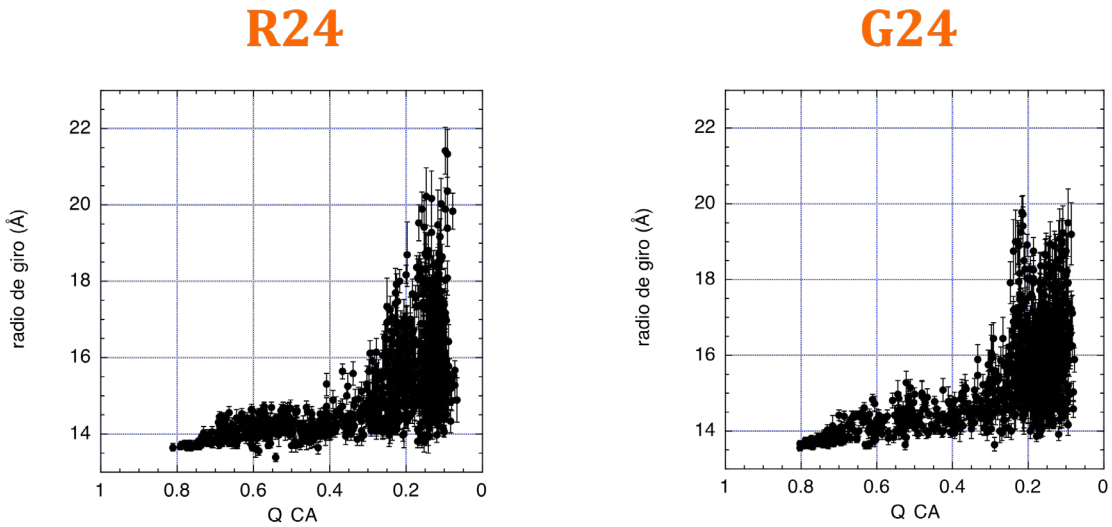
Al aumentar la temperatura a 498K obtenemos la desnaturalización total de ambas proteínas. Una manera de cuantificar este proceso es contabilizar el porcentaje de contactos nativos en función del tiempo de simulación. A continuación tenemos dos gráficas, una por proteína, que muestra los datos de las diez simulaciones distintas. Cada simulación cuenta una historia diferente. Lo más notable es que ninguna trayectoria corresponde a un decaimiento exponencial, y que hay mesetas. Éstas corresponden a mínimos locales en la superficie de energía libre, es decir, a intermediarios de plegamiento.



La cantidad de contactos nativos es fácil de calcular, pero no corresponde a nada que se pueda medir directamente con un experimento. Para poder establecer una conexión con datos experimentales, necesitamos una observable que sí se pueda medir.

Podemos medir el tamaño de la proteína con el radio de giro, cantidad que dice qué tan lejos del centro de masa se encuentran los átomos. Sabemos que por definición el estado nativo es máximamente compacto, y que por lo tanto la desnaturalización implica un aumento en el volumen ocupado por la molécula. Experimentalmente esto se mide con el

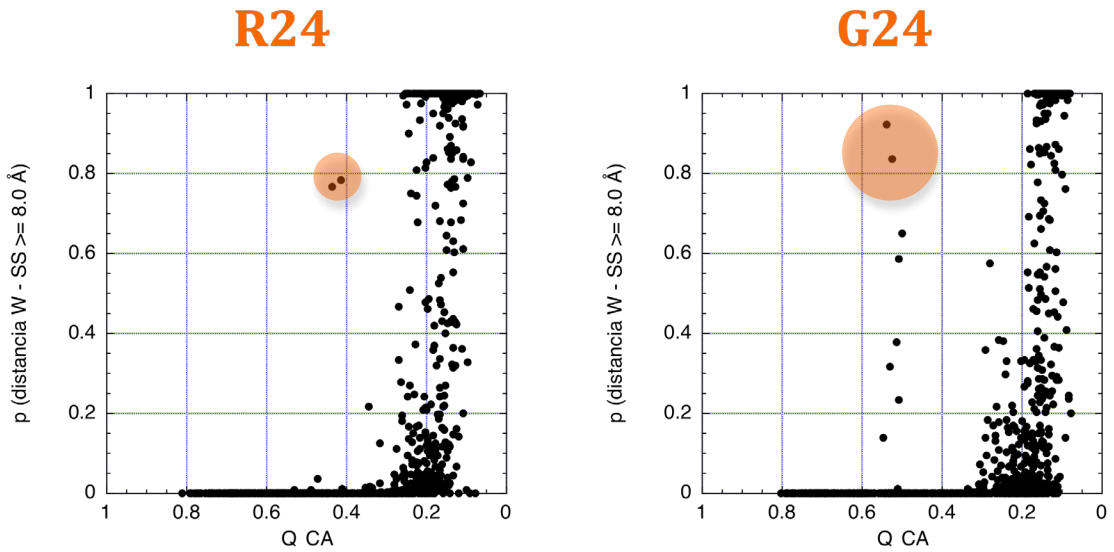
tiempo que tarda la proteína en pasar por una columna densamente empacada con esferas porosas de un tamaño conocido. Mientras más grande es la proteína, más rápidamente sale de la columna. Al desnaturalizar con urea, Blancas y colaboradores (Blancas-Mejía LM *et al.* (2009) J. Mol. Biol. 386:1153) encuentran que el radio de Stokes se incrementa de ~12.5 a casi 27 Å. Como puede verse en las siguientes gráficas, nosotros logramos que la proteína se estire, pero nos quedamos cortos. Esto puede deberse a que no hemos incluido urea en la simulación, y ya se sabe que los ensambles desnaturalizados por temperatura son más compactos que los generados en presencia de urea.



En estas gráficas de radio de giro en función de la cantidad de contactos nativos (Q) vemos que la proteína se va "hinchando" casi linealmente hasta que pierde poco más del 70% de sus contactos nativos, y luego "explota". Esto nos sugiere que la barrera energética principal que separa a la cuenca nativa del ensamble desnaturalizado se encuentra alrededor de Q – 0.3. Otra cosa notable es que **G24** puebla estados con Q ~0.25 y radios de giro cercanos al máximo, que no existen en **R24**. En estos estados pudieran residir conformaciones pro-fibra que expliquen la facilidad de **G24** por iniciar y propagar fibras.

Estas proteínas tienen un solo triptófano, localizado adyacente al único puente disulfuro. Esto es interesante porque el triptófano es un fluoróforo, cuyo espectro de emisión depende fuertemente del ambiente y de la presencia de apagadores como los puentes disulfuro. En el estado nativo, el triptófano está enterrado en el corazón de la proteína, y por su contacto con el puente disulfuro no emite fluorescencia. Al desnaturalizarse la proteína, el triptófano se expone al solvente, resultando en un desplazamiento de Stokes hacia el rojo en el espectro de emisión (el solvente estabiliza al estado excitado, de manera que el fotón emitido es de menor energía que cuando el fluoróforo no está expuesto al solvente); además, se aleja del puente disulfuro lo suficiente como para evitar ser apagado. No hay una distancia clara reportada a la cual esto ocurre, pero se acepta que a distancias mayores a 8Å ya no hay un efecto de apagamiento importante. En los experimentos se interpreta la intensidad de la fluorescencia como una combinación lineal de la emisión producida por el ensamble nativo y la emitida por el ensamble desnaturalizado, sin considerar la presencia de intermediarios de plegamiento (esquema de dos estados). En las siguientes gráficas se muestra la probabilidad de encontrar al

triptófano al menos a 8Å del puente disulfuro, en función de la cantidad de contactos nativos.



Estas dos proteínas prácticamente tienen que estar completamente desnaturalizadas (haber perdido cerca del 80% de los contactos nativos) para que el triptófano emita fluorescencia; las zonas sombreadas marcan conformaciones con un ~45% de contactos nativos que sí serían fluorescentes, que ocurren a muy baja frecuencia. Esto además indica que los intermediarios de plegamiento que vimos en las rutas de desnaturalización son silenciosos espectroscópicamente, reconciliando el comportamiento de dos estados visto en los experimentos con los intermediarios propuestos de las simulaciones.

Otra propiedad espectroscópica que se explota para caracterizar el plegamiento de proteínas es el dicroísmo circular. En esta técnica, se mide la eficiencia de absorción de luz circularmente polarizada hacia la derecha y hacia la izquierda; si hay absorción preferencial de alguna de éstas, se concluye que los cromóforos que absorben esta luz están orientados preferencialmente en el espacio. Si la medición se hace en el ultravioleta cercano (entre 270 y 300 nm), los grupos medidos son las cadenas laterales de los aminoácidos aromáticos tirosina y triptófano. Estas proteínas tienen seis tirosinas y un triptófano. Contamos cuántos estados rotacionales distintos (rotámeros) visita cada una de estas tirosinas y el triptófano a las tres temperaturas simuladas, para determinar cuáles son las que contribuyen al espectro de dicroísmo circular. Cinco de las seis tirosinas están en la superficie de la proteína, por lo que giran libremente aún a baja temperatura. Proponemos que son la tirosina 89 y el triptófano 36 los que dan cuenta de todo el espectro de dicroísmo circular, y para probar esto, la tirosina 89 ya está siendo mutada a alanina. De estar en lo cierto, esta mutación debería de bajar la intensidad de la banda de dicroísmo a la mitad, ya que la intensidad de una banda de absorción depende del número de cromóforos que contribuyen a ella.

Propuesta de rutas de desnaturalización

Habiendo identificado posibles intermediarios de desplegamiento, es muy tentador establecer relaciones temporales entre ellos, armando de esta manera posibles rutas de

desnaturalización. Para ambas proteínas encontramos que los estados iniciales de desnaturalización involucran la separación del asa C'-C" del cuerpo de la proteína. Posteriormente se pierde estructura en la región de CDR1, y en el amino y carboxilo terminal. Recordemos que muchas de estas zonas son proclives a formar fibras, o protegen a zonas que lo son. Al llegar a Q ~ 0.4, la diversidad de estructuras es muy grande, pero tienen en común la presencia de muchas hebras β. Esto es particularmente interesante porque se ha propuesto que las proteínas que se desnaturalizan a estados sin estructura secundaria residual no son buenos formadores de fibras amiloides (Nowak M (2004) Proteins 55:11). Lo que resta por hacer es identificar la estructura mínima necesaria para ser una semilla de fibra amiloide, para cuantificar su presencia en los ensambles desnaturalizados de ambas proteínas, y ver si su abundancia correlaciona con la facilidad para iniciar y/o propagar fibras.

Agradecimientos
Este es un proyecto que hace uso de muchos recursos de supercómputo: KanBalam en DGTIC-UNAM, Argentum en el Centro Nacional de Supercómputo en el IPICyT (San Luis Potosí), Sputnik II en el IBT-UNAM, Entalpia en la FM-UNAM, y Orion en la FC y CIQ-UAEM. Es parte de proyectos financiados por el CONACYT (102182 y 133294). Agradezco a Jessica Araujo, Liliana Martínez, Diana Valenzo, Ángel Santiago, Darely Gutiérrez y David Villaseñor por el análisis bioinformático de las proteínas de pacientes con AL, esencial para el análisis e interpretación de las rutas de desnaturalización.

Propiedades macroscópicas y propagación fotónica en metamateriales

José Samuel Pérez Huerta[1,2], Bernardo Mendoza[3], Guillermo Ortiz[4], W. Luis Mochán Backal[1]

[1]Instituto de Ciencias Físicas, UNAM, Cuernavaca, Mor., México
[2]Posgrado en Ciencias Físicas, UNAM
[3]Centro de Investigaciones en Óptica, León, Gto.
[4] Univ. Nal. Nordeste, Corrientes, Argentina

1 de agosto de 2012

Resumen

Se desarrolla un formalismo extremadamente eficiente para calcular las propiedades dieléctricas macroscópicas de meta-materiales hechos de inclusiones embebidas en una matriz. La geometría de las inclusiones es arbitraria y puede obtenerse de fotografías o dibujos digitalizados que pueden ser manipulados posteriormente con herramientas para procesamiento de imágenes. La composición de las inclusiones y de la matriz es también arbitraria y puede corresponder a aislantes o conductores, con o sin disipación y/o dispersión. En el límite de longitud de onda larga, las dependencias en composición y frecuencia se desacoplan, lo cual permite una aceleración del cálculo por varios órdenes de magnitud frente a métodos alternativos. Como una aplicación, diseñamos un sistema extremadamente anisotrópico que se comporta como un absorbedor o un reflector casi perfecto sobre un rango de frecuencias entonable de acuerdo a la dirección de polarización, y obtenemos una explicación a la transmitancia extraordinaria de películas metálicas perforadas por huecos nanométricos, sin invocar para ello a la propagación de plasmones de superficie. Cuando la longitud de onda es comparable con la celda unitaria, la respuesta macroscópica adquiere una fuerte dispersión espacial. Al tomarla en cuenta, nuestro formalismo macroscópico es capaz de reproducir las bandas fotónicas del sistema.

1. Introducción

Consideremos un sistema como el ilustrado esquemáticamente en la figura 1, fabricado con partículas de cierto material inmersas en un segundo material. Si

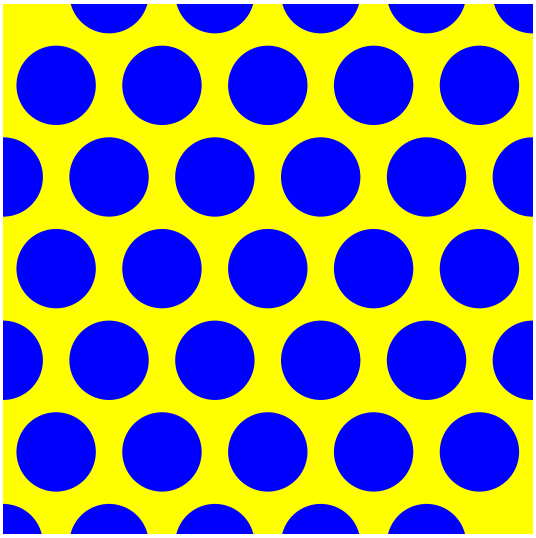


Figura 1: Cristal fotónico formado por partículas de un material inmersas periódicamente en una matriz de otro material.

el sistema se repitiera periódicamente, podríamos considerarlo un cristal *artificial*. Así como en un cristal real la propagación de excitaciones tales y como los portadores de carga, electrones y huecos, o los cuantos de vibración, fonones, se describe en términos de bandas permitidas con ciertas *relaciones de dispersión* complicada y con ciertas brechas prohibidas, la propagación del campo electromagnético en cristales artificiales también podría describirse en términos de bandas y brechas (fig. 2). Esencialmente, en cada uno de los materiales la luz se propaga como ondas planas libres, pero éstas *chocan* y se esparcen al llegar a la frontera que divide los dos materiales. Las ondas esparcidas por distintas inclusiones pueden interferir constructivamente en la dirección opuesta a su dirección original de propagación cuando su longitud de onda es del orden de la celda unitaria del cristal, sufriendo una *reflexión de Bragg* que les impide continuar su propagación en la dirección original. Estas reflexiones son las responsables de abrir brechas en la relación de dispersión, i.e., en la gráfica de la frecuencia ω vs. el vector de onda $\vec{k}$. Los detalles de la relación de dispersión, el ancho y frecuencia de la brecha y su dependencia en la dirección de propagación, la existencia de una brecha absoluta en que la luz no pueda propagarse en *ninguna* dirección y de manera más general, todas las propiedades electromagnéticas de estos *cristales fotónicos*, dependen de la composición del mismo, es decir, de los materiales de que están compuestas las inclusiones y el anfitrión. Sin embargo, dependen además de la geometría de la red cristalina y de la geometría de cada una de las partículas que conforman al medio.

Consideremos ahora un medio fabricado con placas aislantes sobre las cuales se han *dibujado* pistas conductoras, como se hace en los circuitos impresos. En

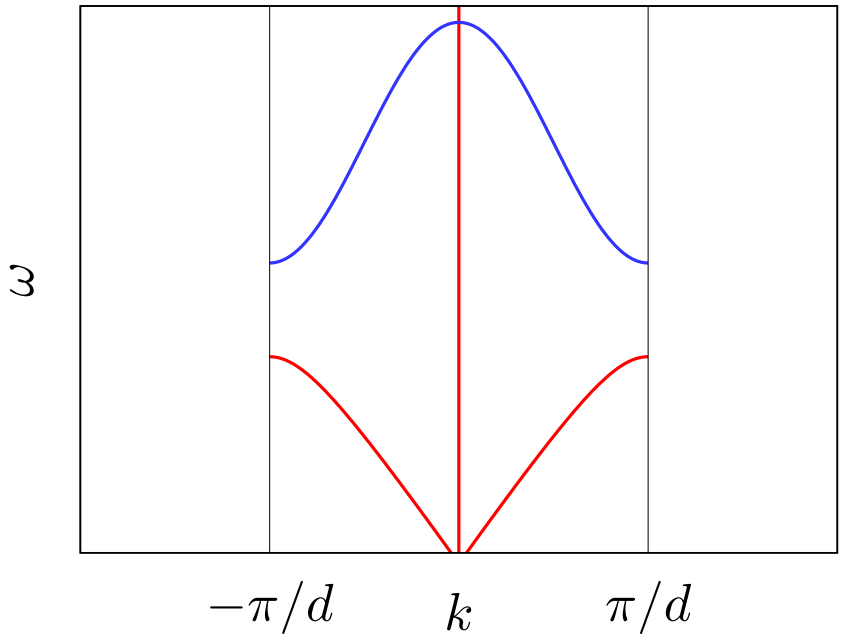


Figura 2: Relación de dispersión esquemática para la propagación del campo electromagnético en el seno de un cristal artificial. Se muestra una banda *acústica* (rojo) y una banda *óptica* (azul) y se indica el borde de la zona de Brillouin.

particular, consideremos pistas en forma de un círculo conductor interrumpido y terminado con dos líneas paralelas, como ilustramos en la fig. 3. Frente a un campo electromagnético, cada círculo conductor se comporta como una pequeña inductancia y cada pareja de líneas se comporta como un pequeño capacitor, de forma que la figura muestra un arreglo de circuitos LC acoplados entre sí. Un campo electromagnético puede inducir un dipolo magnético en cada espira y un dipolo eléctrico en cada capacitor. Escogiendo adecuadamente los parámetros del sistema, podríamos lograr una resonancia en la respuesta magnética y otra en la respuesta dieléctrica. Típicamente, las funciones respuesta cambian de signo arriba de una resonancia, por lo cual, podríamos construir materiales con permitividad *efectiva* ϵ y permeabilidad efectiva μ negativas (fig. 4).

Recordemos que la relación de dispersión de una onda electromagnética es $k^2 = \epsilon\mu\omega^2/c^2$. Por ello, k es una cantidad imaginaria cuando $\epsilon < 0$ o $\mu < 0$, por lo cual el campo no se puede propagar en esas condiciones y las ondas electromagnéticas decaen exponencialmente en una distancia del orden de $\delta = 1/\mathrm{Im}k$. Sin embargo, si ambas cantidades son negativas *simultáneamente*, k vuelve a ser un número real y el campo electromagnético sí se puede propagar. Ahora, recordemos que el campo eléctrico $\vec{E}$, la densidad de flujo magnético $\vec{B}$ y el vector de onda $\vec{k}$ forman una triada ordenada derecha, mientras que $\vec{E}$, el campo magnético $\vec{H}$ y el la densidad de flujo de energía $\vec{S}$ también. Como $\vec{H}$ y $\vec{B}$ apuntan en direcciones opuestas cuando $\mu < 0$, entonces $\vec{S}$ y $\vec{k}$ también. Esto lleva al curioso resultado de tener ondas cuya fase se propaga en una dirección mientras que *su energía se propaga en la dirección opuesta.* Una consecuencia de esto es que cuando la luz penetra en uno de estos materiales artificiales conocidos como *medios izquierdos*, se refracta de acuerdo a la *ley de Snell* pero con un índice de refracción negativo; el haz refractado se halla del lado opuesto de la normal a la superficie de lo que sucede en un medio normal, lo cual permite en principio crear lentes planas perfectas (fig. 5). El rango de frecuencias en las que el material presenta este comportamiento curioso depende de las propiedades de los materiales aislantes y conductores, pero también depende de la geometría de los anillos cortados.

Finalmente consideremos una película delgada metálica iluminada por un haz luminoso cuya frecuencia sea menor a su frecuencia de plasma (fig. 6). La película sería opaca y lo seguiría siendo aún si se le practicara una serie de agujeros, siempre y cuando éstos tuvieran un diámetro mucho menor que la longitud de onda, pues la luz *no cabría* a través de ellos. Sin embargo, se han observado transmitancias varios órdenes mayores a la esperado en estos materiales [1]. Una explicación tentativa ha sido la excitación de plasmones de superficie mediados por la textura de la película [2]. Los plasmones de superficie son excitaciones que se propagan sobre la superficie de metales y cuyo campo eléctrico decae exponencialmente con la distancia a la superficie. Tienen la peculiaridad de no poder excitarse por luz en una superficie plana, pero si pueden excitarse en una superficie rugosa o con algún tipo de textura que esparza la luz, y su longitud de onda puede ser mucho menor que la de la luz de la misma frecuencia. Por ello, los plasmones de superficie si podrían atravesar por agujeros delgados y así llegar al otro lado de la película, donde podrían volverse a esparcir convirtiéndose

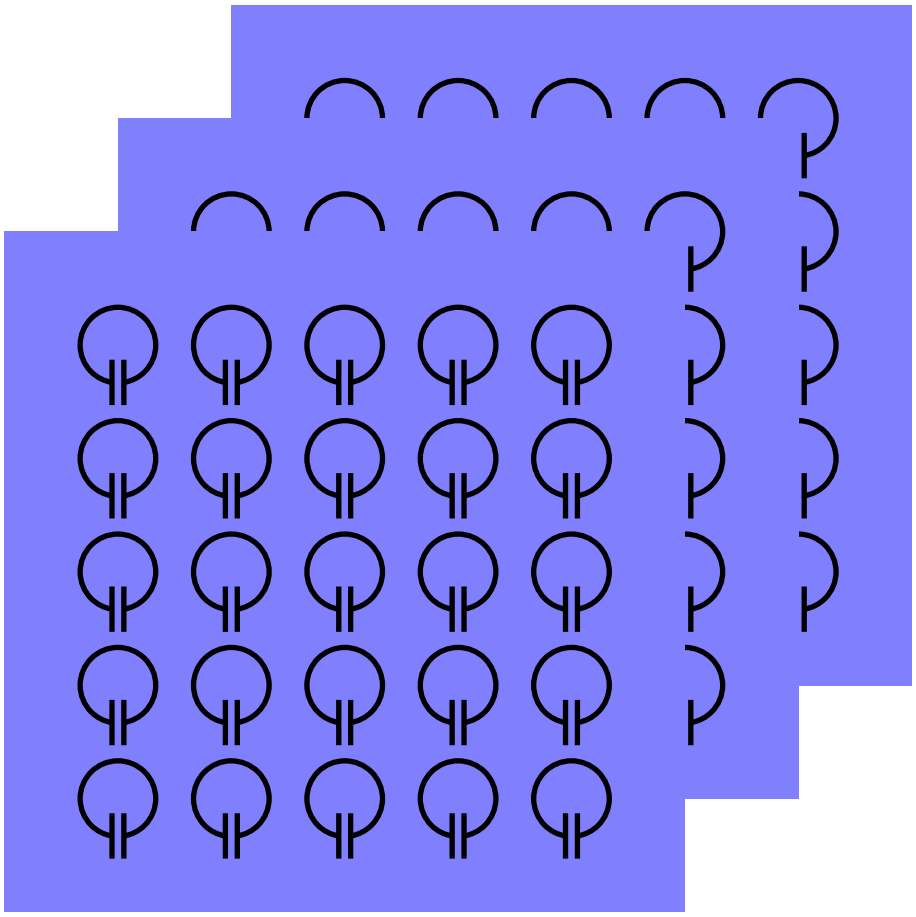


Figura 3: Se muestra un arreglo de círculos conductores interrumpidos y terminados en un par de líneas paralelas dibujado sobre matrices aislantes.

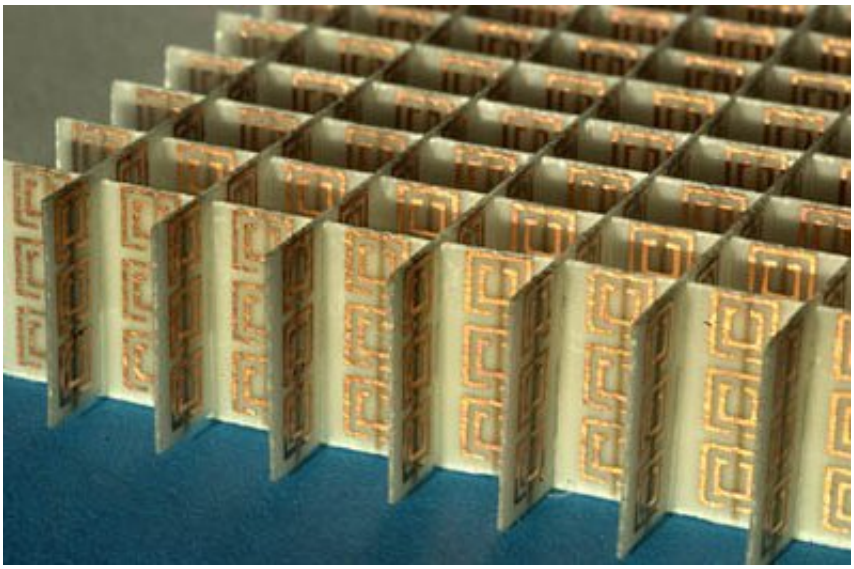


Figura 4: Realización de un material izquierdo con índice de refracción negativo en la región de microondas.

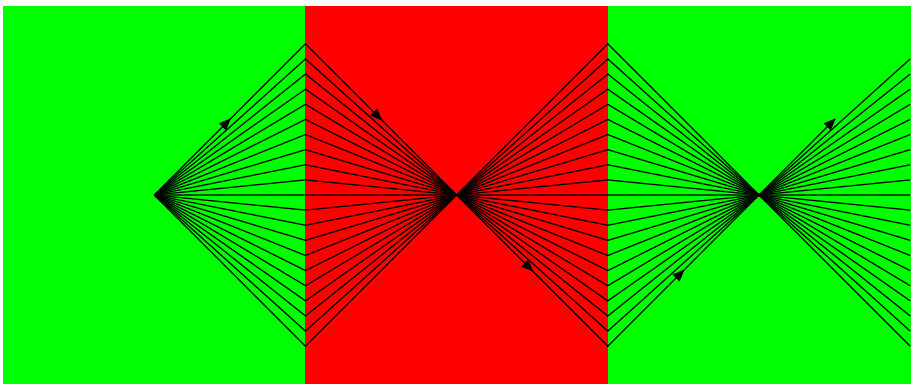


Figura 5: Película delgada plana de una material izquierdo que funciona como lente perfecta. La luz que viaja hacia arriba (abajo) se refracta hacia abajo (arriba) al entrar o salir de la película.

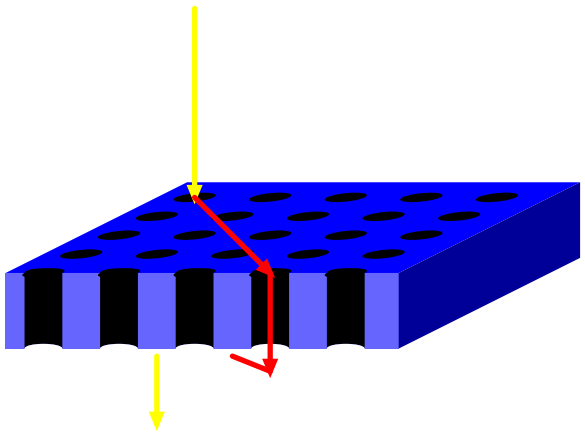


Figura 6: Película metálica delgada con un arreglo de agujeros cilíndricos. Se muestra un haz de luz incidente (flechas amarillas) que se acopla a un plasmón de superficie (flechas rojas) capaz de atravesar los agujeros y convertirse en luz del otro lado de la película.

en el haz de luz transmitido.

En los ejemplos anteriores hemos descrito algunos de los muchos sistemas cuyas propiedades electromagnéticas están determinadas no solamente por su composición sino también por su geometría, definida por su tamaño, forma y repetición. El problema al que nos queremos enfrentar es al del cálculo de estas propiedades. En el resto de este artículo desarrollaremos un formalismo que permite el cálculo eficiente de la respuesta dieléctrica de sistemas periódicos de geometría y composición arbitraria.

2. Teoría

2.1. Límite de Longitud de Onda Larga

Considere un sistema con cierta *textura* espacial y descrito por cierta respuesta dieléctrica $\hat{\epsilon}$, definida mediante la ecuación

$$\vec{D} = \hat{\epsilon}\vec{E}, \tag{1}$$

donde $\vec{D}$ es el desplazamiento y $\vec{E}$ el campo eléctrico. Escribimos $\hat{\epsilon}$ de manera general como un operador lineal. Así, dejamos implícita su dependencia en la frecuencia, equivalente a una dependencia en el tiempo de retraso, y nos permitimos entender la Ec. (1) como una ecuación en el dominio de la frecuencia, donde ϵ es una simple *constante* multiplicativa compleja para cada frecuencia, o como una ecuación en el dominio del tiempo, donde $\hat{\epsilon}$ es un operador integral

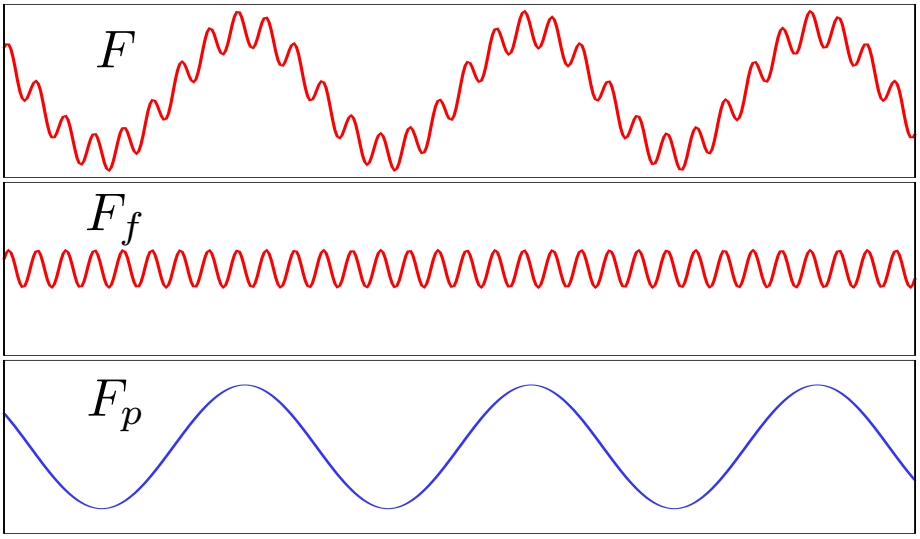


Figura 7: Un campo típico F en el interior de un meta-material (panel superior). Debido a la textura del material, el campo tiene una parte que oscila rápidamente (F_f panel central) además de una parte que oscila lentamente en el espacio (F_p panel inferior).

cuyo kernel $\epsilon(t-t')$ toma en cuenta que la polarización al tiempo t depende del campo eléctrico no sólo al tiempo t sino también a tiempos previos $t' < t$. La dependencia en ω o en $t-t'$ es conocida como *dispersión temporal*. Asimismo, dejamos implícita la dependencia en la posición $\vec{r}$ en que evaluamos el desplazamiento y en la posición $\vec{r}'$ en que actúa el campo eléctrico. En materiales sin *dispersión espacial* bastaría considerar el caso $\vec{r} = \vec{r}'$.

En un sistema como los descritos en la sección 1 todos los campos se verían esquemáticamente como el ilustrado en la figura 7. El campo tendría fluctuaciones espaciales en una escala de distancia d pequeña determinada por las escalas de distancia de la textura, tales y como el tamaño de las partículas inmersas en la matriz y la distancia entre ellas, así como oscilaciones en una escala de distancias grande determinada por la longitud de onda libre caracterizada por la escala de variación temporal, i.e., por la distancia $\lambda = cT$ que recorre el campo durante un periodo T a la velocidad de la luz c. Esto sugiere definir dos operadores, el proyector *promedio* $\hat{P}_p$ y el proyector *fluctuaciones* $\hat{P}_f$, tales que para cualquier campo F podamos escribir su promedio como $F_p = \hat{P}_p F$ y sus fluctuaciones como $F_f = \hat{P}_f F$. Necesariamente, $\hat{P}_p$ y $\hat{P}_f$ deben ser idempotentes, su producto mutuo es nulo y su suma es la identidad. Aplicando estos proyectores a la ec. (1) podemos escribir

$$\vec{D}_\alpha = \sum_\beta \hat{\epsilon}_{\alpha\beta} \vec{E}_\beta. \tag{2}$$

Aquí hemos definido los *elementos de matriz* $\hat{\epsilon}_{\alpha\beta} \equiv \hat{P}_\alpha \hat{\epsilon} \hat{P}_\beta$ con $\alpha, \beta = p, f$. Podemos identificar a los campos macroscópicos con los campos promediados, por lo que la respuesta dieléctrica macroscópica se define mediante

$$\vec{D}_p = \hat{\epsilon}_M \vec{E}_p. \tag{3}$$

Comparando las ecs. (3)con (2) vemos que $\hat{\epsilon}_M = \hat{\epsilon}_{pp} + \ldots$, i.e., la respuesta macroscópica es el promedio de la respuesta microscópica más una corrección, la cual se conoce como *efecto de campo local* y proviene del acoplamiento entre campos promedio y campos fluctuantes.

Recordemos ahora que el teorema de Helmholtz nos permite descomponer todo campo vectorial $\vec{F} = \vec{F}^L + \vec{F}^T$ en una parte longitudinal $\vec{F}^L = \hat{P}^L \vec{F}$ y una transversal $\vec{F}^T = \hat{P}^T \vec{F}$, tales que $\nabla \times F^L \equiv 0$ y $\nabla \cdot F^T \equiv 0$, de manera que $\nabla \cdot F^L = \nabla \cdot F$ y $\nabla \times F^T = \nabla \times F$. Formalmente, realizamos esta separación empleando los proyectores $\hat{P}^L = \nabla \nabla^{-2} \nabla$ y $\hat{P}^T = \hat{1} - \hat{P}^L$, donde definimos ∇^{-2} como el *inverso* del operador Laplaciano, $\nabla^2 \nabla^{-2} = \hat{1}$, el cual podemos identificar con un operador integral cuyo Kernel es la función de Green para la ecuación de Poisson $-1/(4\pi|\vec{r} - \vec{r}'|)$. Introduciendo múltiples instancias de la identidad $\hat{1} = \hat{P}^L + \hat{P}^T$ en la ecuación (2), podemos reescribirla como

$$\vec{D}^I = \sum_J \hat{\epsilon}^{IJ} \vec{E}^J, \tag{4}$$

donde $\hat{\epsilon}^{IJ} = \hat{P}^I \hat{\epsilon} \hat{P}^J$.

En el *límite no retardado*, en el cual la escala interna de distancia d es mucho menor que la longitud de onda λ, podemos obtener la respuesta macroscópica del sistema notando que la parte *longitudinal* del desplazamiento $\vec{D}^L$ es la parte longitudinal del *campo eléctrico externo* $\vec{E}_{ext}$; tiene el mismo rotacional $\nabla \times \vec{D}^L = \nabla \times \vec{E}^L_{ext} = 0$ y las mismas fuentes $\nabla \cdot \vec{D}^L = \nabla \cdot \vec{E}^L_{ext} = 4\pi\rho_{ext}$, donde ρ_{ext} es la densidad de carga externa. En el límite no retardado la luz recorre una distancia del orden de d en un tiempo despreciable comparable con su periodo, como si su velocidad de propagación fuese prácticamente infinita. Por ello, el campo eléctrico puede obtenerse en dicho límite empleando una aproximación cuasi-estática, en la que tomamos el límite $c \to \infty$ en las ecuaciones de Maxwell. En dicho límite, el campo eléctrico se puede derivar de un potencial escalar $\vec{E} = -\nabla\phi$ y es por tanto longitudinal, i.e., $\vec{E} = \vec{E}^L$. En este caso $\vec{D}^L = \hat{\epsilon}^{LL} \vec{E}^L$, de donde

$$\vec{E}^L = (\hat{\epsilon}^{LL})^{-1} \vec{D}^L. \tag{5}$$

Aplicando $\hat{P}_p$ a esta ecuación para promediarla,

$$\vec{E}^L_p = (\hat{\epsilon}^{LL})^{-1}_{pp} \vec{D}^L_p + (\hat{\epsilon}^{LL})^{-1}_{pf} \vec{D}^L_f, \tag{6}$$

pero, como $\vec{D}^L$ es el campo externo, entonces *no tiene fluctuaciones* inducidas por la textura del sistema. Si el campo externo tuviera fluctuaciones, no tendría sentido buscar una respuesta macroscópica. Eso significa que podemos eliminar el segundo término del lado derecho de la ec. (5) para obtener

$$\vec{E}^L_p = (\hat{\epsilon}^{LL})^{-1}_{pp} \vec{D}^L_p. \tag{7}$$

Finalmente, notamos que ambos lados de la ec. (7) involucran campos macroscópicos, por lo cual podemos identificar la respuesta macroscópica

$$(\hat{\epsilon}_M^{LL})^{-1} = (\hat{\epsilon}^{LL})_{pp}^{-1}. \tag{8}$$

Este resultado nos dice que en el límite no retardado, *el inverso de la proyección longitudinal de la respuesta dieléctrica macroscópica es el promedio del inverso de la parte longitudinal de la respuesta dieléctrica microscópica*. Este resultado es un caso particular de un enunciado más general: *la respuesta macroscópica a una excitación externa es el promedio de la correspondiente respuesta microscópica*.

2.2. Sistema periódico

Consideremos ahora un sistema periódico caracterizado por una red de Bravais $\{\vec{R} = \sum_{i=1}^{D} n_i \vec{a}_i\}$ donde n_i son D números enteros, $\vec{a}_i$ son D vectores primitivos y $D = 1, 2$ o 3 es la dimensionalidad del sistema. Los vectores $\vec{R}$ son tales que el sistema es idéntico si lo vemos desde una posición $\vec{r}$ o desde cualquier otra posición de la forma $\vec{r} + \vec{R}$, es decir, el sistema no cambia si damos cualquier cantidad de pasos (enteros) en la dirección y del tamaño del vector $\vec{a}_i$.

Asociada a la red *directa* $\{\vec{R}\}$, se define una *red recíproca* $\{\vec{K}\}$ formada por todos los vectores tales que $\exp(i\vec{K} \cdot \vec{R}) = 1$. El teorema de *Bloch* implica que todo campo en el interior del cristal, y en particular, el campo eléctrico $\vec{E}(\vec{r})$, pueden escribirse en términos de eigen-funciones del operador de translación $\hat{T}_{\vec{R}}$ por vectores $\vec{R}$, de ondas de Bloch de la forma

$$\vec{E}_{\vec{k}}(\vec{r}) = \sum_{\vec{K}} \vec{E}_{\vec{k},\vec{K}} e^{i(\vec{k}+\vec{K})\cdot\vec{r}}, \tag{9}$$

donde el eigen-valor de $T_{\vec{R}}$ es $e^{i\vec{k}\cdot\vec{R}}$. El vector $\vec{k}$ se conoce como el vector de Bloch y típicamente, pero no necesariamente se escoge en la primera zona de Brillouin, i.e., de manera que k sea más pequeño que $|\vec{k} + \vec{K}|$ para todo vector recíproco $\vec{K}$. En la ec. (9) escribimos la onda de Bloch de manera explícita como una suma de ondas planas, pues los operadores diferenciales toman una forma especialmente simple, $\nabla \to i(\vec{k} + \vec{K})$, al actuar sobre ellas. Además tomaremos ondas monocromáticas con frecuencia ω. Notamos que la interacción de una onda plana con la red periódica simplemente produce una difracción con vector de difracción $\vec{K}$ de la forma $\vec{q}+\vec{K}' \to \vec{q}+\vec{K}''$ con $\vec{K} = \vec{K}'' - \vec{K}'$, y que la frecuencia se conserva en este proceso. En el caso $k \ll K$ para todos los vectores recíprocos $\vec{K} \neq 0$ es razonable identificar a las ondas planas con vector de onda $\vec{k}+\vec{K}$ como fluctuaciones y al campo con vector de onda $\vec{k}$ como el campo promedio. De esta forma, podemos identificar al proyector promedio como aquel que elimina del campo las contribuciones con vector recíproco no nulo $\hat{P} \to \delta_{\vec{K}0}$. Por otro lado, el proyector longitudinal es aquel que proyecta sobre la dirección del vector de onda, i.e.,

$$\hat{P}^L \to P^L_{\vec{K}\vec{K}'} = \frac{\vec{k}+\vec{K}}{|\vec{k}+\vec{K}|}\frac{\vec{k}+\vec{K}}{|\vec{k}+\vec{K}|} = \hat{K}\hat{K}, \tag{10}$$

donde definimos $\hat{K} = (\vec{k} + \vec{K})/|\vec{k} + \vec{K}| \approx \vec{K}/K$ y donde la aproximación es válida cuando el vector de Bloch es muy cercano a cero y $\vec{K} \neq 0$.

Representando al desplazamiento y al campo eléctrico en términos de sus componentes $\vec{D}_{\vec{K}}$ y $\vec{E}_{\vec{K}}$ (omitiendo a la cantidad conservada $\vec{q}$ para simplificar la notación), podemos reescribir la ec. (1) como

$$\vec{D}_{\vec{K}} = \sum_{\vec{K}'} \epsilon_{\vec{K}\vec{K}'} \vec{E}_{\vec{K}'}, \tag{11}$$

donde $\epsilon_{\vec{K}\vec{K}'}$ es la transformada de Fourier de la respuesta $\epsilon(\vec{r})$ con vector de onda $\vec{K} - \vec{K}'$. La *proyección* longitudinal de $\hat{\epsilon}$ queda representada entonces por

$$\epsilon^{LL}_{\vec{K}\vec{K}'} = \hat{K} \eta_{\vec{K}\vec{K}'} \hat{K}', \tag{12}$$

donde introdujimos la *componente* longitudinal de la respuesta

$$\eta_{\vec{K}\vec{K}'} = \hat{K} \cdot (\epsilon_{\vec{K}\vec{K}'} \hat{K}'). \tag{13}$$

La inversa de la respuesta longitudinal, calculada sobre el sub-espacio de los campos longitudinales es entonces

$$(\epsilon^{LL})^{-1}_{\vec{K}\vec{K}'} = \hat{K} \eta^{-1}_{\vec{K}\vec{K}'} \hat{K}', \tag{14}$$

como puede verificarse mediante una simple multiplicación,

$$\begin{aligned} \sum_{\vec{K}''} (\epsilon^{LL})^{-1}_{\vec{K}\vec{K}''} (\epsilon^{LL})_{\vec{K}''\vec{K}'} &= \sum_{\vec{K}''} \hat{K} \eta^{-1}_{\vec{K}\vec{K}''} \hat{K}'' \cdot \hat{K}'' \eta_{\vec{K}''\vec{K}'} \hat{K}' & (15) \\ &= \sum_{\vec{K}''} \hat{K} \eta^{-1}_{\vec{K}\vec{K}''} \eta_{\vec{K}''\vec{K}'} \hat{K}' = \sum_{\vec{K}''} \hat{K} \delta_{\vec{K}\vec{K}'} \hat{K}' & (16) \\ &= \hat{K}\hat{K}' = P^{L}_{\vec{K}\vec{K}'}; & (17) \end{aligned}$$

recordemos que el proyector longitudinal es como el operador identidad cuando actúa sobre campos longitudinales.

Regresando a la ecuación (8), la podemos escribir como

$$(\epsilon^{LL}_M)^{-1} = \hat{k} \eta^{-1}_{00} \hat{k}, \tag{18}$$

en términos de el vector unitario $\hat{k} = \vec{k}/k$ y el elemento 00 de la inversa de la matriz $\eta_{\vec{K}\vec{K}'}$.

En resumen, para obtener la respuesta dieléctrica macroscópica tenemos que calcular la componente longitudinal $\eta_{\vec{K}\vec{K}'}$ de la respuesta dieléctrica microscópica, invertirla como matriz con índices $\vec{K}$ y $\vec{K}'$ y finalmente tomar la componente 00 de dicha inversa. Este procedimiento nos lleva a la proyección de ϵ^{-1}_M a lo largo del vector de onda (Bloch) $\vec{k}$. Tomando el límite $\vec{k} \to 0$ a lo largo de varias direcciones $\hat{k}$ podemos obtener todas las componentes de ϵ^{-1}_M y una última inversión nos lleva a el *tensor* dieléctrico macroscópico en el límite de longitud de onda larga.

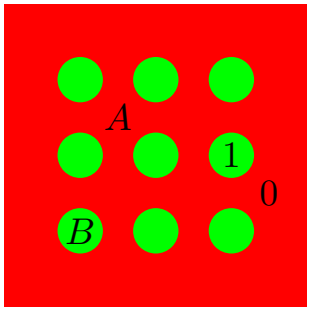


Figura 8: Sistema de dos componentes formado por inclusiones B dispuestas periódicamente en una matriz A. Se indican las funciones dieléctricas y los valores de la función característica.

2.3. Sistema de dos componentes

A continuación desarrollaremos la teoría anterior al caso de un sistema formado por inclusiones de un material B colocadas de manera periódica dentro de una matriz A, suponiendo que ambos materiales tienen un funciones dieléctricas ϵ_A y ϵ_B bien definidas (fig. 8). En este caso la respuesta microscópica es

$$\epsilon(\vec{r}) = \begin{cases} \epsilon_B & \text{si } \vec{r} \text{ en alguna inclusión } B \\ \epsilon_A & \text{si } \vec{r} \text{ en la matriz } A \end{cases} \tag{19}$$

y la podemos escribir como

$$\epsilon(\vec{r}) = \epsilon_{AB}(u - B(\vec{r})), \tag{20}$$

donde $\epsilon_{AB} \equiv \epsilon_A - \epsilon_B$ y $u = \epsilon_A/\epsilon_{AB}$ es la llamada *variable espectral*. De acuerdo a la ec. (18), debemos calcular el elemento 00 de

$$(\eta_{\vec{K}\vec{K}'})^{-1} = \frac{1}{\epsilon_{AB}}(u - B^{LL}_{\vec{K}\vec{K}'})^{-1}. \tag{21}$$

Este problema es análogo al del cálculo cuántico de la función de Green proyectada

$$\mathcal{G}_{00}(\varepsilon) = \langle 0|\hat{\mathcal{G}}(\varepsilon)|0\rangle \tag{22}$$

sobre cierto estado $|0\rangle$, donde $\hat{\mathcal{G}}(\varepsilon) = \varepsilon - \hat{\mathcal{H}}$ es el operador de Green correspondiente a cierto Hamiltoniano $\mathcal{H}$ evaluada en cierta energía compleja ε [3]. Las ecuaciones (21) y (22) sugieren identificar

$$\hat{\mathcal{H}} \to B^{LL}_{\vec{K}\vec{K}'} \equiv \hat{K} \cdot B(\vec{K} - \vec{K}')\hat{K}', \tag{23}$$

y a la energía ε con la variable espectral u.

2.4. Recursión de Haydock

La identificación previa nos permite tomar prestados métodos inicialmente desarrollados para el cálculo de funciones de Green en mecánica cuántica y adaptarlos al cálculo de la función dieléctrica macroscópica.

Empecemos por identificar al estado $|0\rangle$ como aquel correspondiente a una onda plana con vector recíproco $\vec{K} = 0$. En una base de ondas planas, este estado está representado por la delta de Kronecker $\delta_{\vec{K}0}$. Definamos ahora un estado nulo $|-1\rangle \equiv 0$ y construyamos nuevos estados de acuerdo a la recursión

$$|\widetilde{n+1}\rangle \equiv \hat{\mathcal{H}}|n\rangle = h_{nn+1}|n+1\rangle + \sum_{m\leq n} h_{nm}|m\rangle, \tag{24}$$

donde $|n+1\rangle$ es un nuevo estado a determinar a partir de todos los anteriores $|m\rangle$ y h_{nm} son coeficientes a determinar a partir de la condición de ortogonalidad del nuevo estado con todos los anteriores $\langle m|n+1\rangle = 0$ $(m \leq n)$. Si escogemos coeficientes reales, entonces $h_{mn} = h_{nm}$ debido a la hermiticidad de $\mathcal{H}$. Notamos que si la multiplicamos la ec. (24) por $\langle m|$ con $m \leq n-2$ el resultado es nulo, pues

$$\langle m|\widetilde{n+1}\rangle = \langle m|\mathcal{H}|n\rangle = \langle n|\mathcal{H}|m\rangle, \tag{25}$$

y $\mathcal{H}|m\rangle$ tiene contribuciones $|p\rangle$, con $p \leq m+1 < n$. Por lo tanto, podemos escribir

$$|\widetilde{n+1}\rangle = b_{n+1}|n+1\rangle + a_n|\tilde{n}\rangle + b_n|n-1\rangle. \tag{26}$$

Multiplicando la ec. (24) por $\langle n|$ obtenemos

$$a_n \equiv h_{nn} = \langle n|\widetilde{n+1}\rangle = \langle n|\mathcal{H}|n\rangle. \tag{27}$$

Suponiendo que conocemos el coeficiente b_n a partir de una iteración anterior, solo nos faltaría calcular el coeficiente b_{n+1}, el cual podemos obtener de la condición de normalización

$$\langle n+1|n+1\rangle = 1 \tag{28}$$

de donde obtenemos

$$b_{n+1} = [\langle\widetilde{n+1}|\widetilde{n+1}\rangle - a_{n+1}^2 - b_n^2]^{1/2}. \tag{29}$$

Finalmente, obtenemos una expresión para el nuevo estado

$$|n+1\rangle = \frac{1}{b_{n+1}}\left(|\widetilde{n+1}\rangle - a_n|n\rangle - b_n|n-1\rangle\right). \tag{30}$$

Notamos que iterando las ecuaciones (27), (29), (30) partiendo de $n = 0$ y dando el valor inicial $b_0 = 0$, construimos una base de estados $|n\rangle$ en la cual el hamiltoniano $\mathcal{H}$ está representado por una matriz tri-diagonal

$$\hat{\mathcal{H}} \to (h_{nm}) = \begin{pmatrix} a_0 & b_1 & 0 & 0 & 0 & \cdots \\ b_1 & a_1 & b_2 & 0 & 0 & \cdots \\ 0 & b_2 & a_2 & b_3 & 0 & \cdots \\ 0 & 0 & b_3 & a_3 & b_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{31}$$

Para construir esta matriz es necesario aplicar repetidamente el Hamiltoniano a estados $|n\rangle$ consecutivos. Empleando la ec. (23) notamos que podemos empezar representando al estado como superposición de ondas planas con coeficientes $\langle\vec{K}|n\rangle$ y multiplicando éstos por el vector unitario $\vec{K}$. Luego, en lugar de hacer una convolución con la transformada de Fourier de la función característica $B(\vec{K}-\vec{K}')$ podemos hacer una transformada de Fourier y multiplicar por la función característica $B(\vec{r})$ en espacio real. Finalmente, tomamos la transformada de Fourier inversa y multiplicamos escalar-mente cada elemento con el vector unitario $\hat{K}$. El resultado será el estado $|\widetilde{n+1}\rangle$, obtenido ¡sin haber realizado ninguna operación matricial!

Para proseguir con el cálculo de la respuesta macroscópica necesitamos, de acuerdo a la ec. (21), invertir la matriz tri-diagonal $u\delta_{nm} - h_{nm}$. Sin embargo, no necesitamos la matriz completa, sino únicamente el elemento 00. Dada la estructura tri-diagonal, esto puede hacerse recursivamente. Empezamos por escribir $u\delta_{nm} - h_{nm} \equiv \mathcal{K}_0$, donde definimos recursivamente las matrices

$$\mathcal{K}_n = \left(\begin{array}{c|c} u - a_n & -\mathcal{R}_{n+1} \\ \hline -\mathcal{R}_{n+1}^T & \mathcal{K}_{n+1} \end{array}\right) \tag{32}$$

por bloques, donde $\mathcal{R}_n = (b_{n+1}, 0, 0, \ldots)$ es una matriz renglón y $\mathcal{R}_n^T$ su transpuesta. Escribimos la inversa de $\mathcal{K}_n$ en bloques como

$$\mathcal{K}_n^{-1} = \left(\begin{array}{c|c} s_n & \mathcal{T}_{n+1} \\ \hline \mathcal{T}_{n+1}^T & \mathcal{U}_{n+1} \end{array}\right), \tag{33}$$

donde s_n es un número, $\mathcal{T}_{n+1}$ es un vector renglón, $\mathcal{T}_{n+1}^T$ su transpuesta y $\mathcal{U}_{n+1}$ una matriz cuadrada. A partir de la identidad $\mathcal{K}_n^{-1}\mathcal{K}_n = \mathbf{1}$ es fácil demostrar que $s_n = (u - a_n - \mathcal{R}_{n+1}\mathcal{K}_n^{-1}\mathcal{R}_n^T)^{-1}$, y como $\mathcal{R}_n$ tiene sólo un elemento distinto de cero, esta ecuación se simplifica a $s_n = (u - a_n - b_n^2 s_{n+1})^{-1}$. Finalmente, substituyendo recursivamente s_{n+1} en s_n hasta llegar a s_0, y sustituyendo este resultado en la ec. (18) llegamos al resultado final de esta sección,

$$\hat{k}\cdot\epsilon_M^{-1}\cdot\hat{k} = \frac{1}{\epsilon_{AB}}\cfrac{1}{u - a_0 - \cfrac{b_1^2}{u-a_1-\cfrac{b_2^2}{u-a_2-\cfrac{b_3^2}{\ddots}}}} \tag{34}$$

Repitiendo el cálculo anterior para distintas direcciones del vector de onda $\hat{k}$ podemos obtener todas las componentes de ϵ_M^{-1} y por lo tanto de ϵ_M.

Notamos que en la ec. (34) aparecen los *coeficientes de Haydock* a_n y b_n, los cuales dependen exclusivamente de la geometría del sistema, caracterizada por la red de Brillouin $\{\vec{R}\}$, su red recíproca $\{\vec{K}\}$ y por la función característica $B(\vec{r})$, y no dependen en absoluto de la composición del sistema ni de la frecuencia con que oscila el campo. Los resultados dependen de la frecuencia y de la composición del sistema exclusivamente a través de la variable espectral u. Esto implica que para una geometría podemos calcular una sola vez los coeficientes de

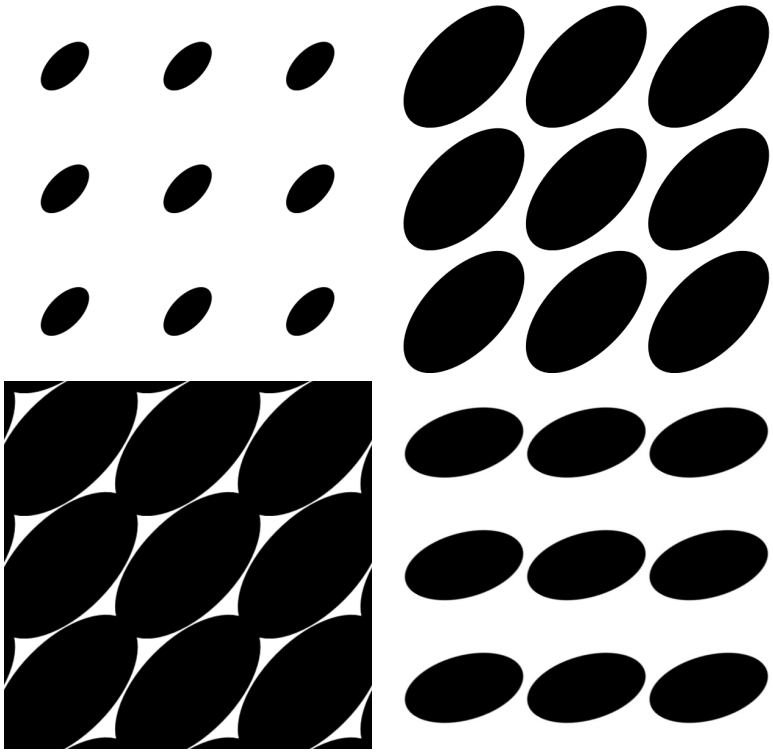


Figura 9: Nuestro cálculo parte de la función característica la cual puede obtenerse de una imagen digitalizada del sistema, misma que puede rotarse y/o escalarse digitalmente. La figura muestra varios meta-materiales obtenidos a partir de una sola imagen mediante escalamientos y rotaciones seguidos de repeticiones periódicas.

Haydock y con ellos hacer cálculos para muchos materiales y calcular espectros completos para amplios rangos de frecuencia simplemente dando valores diversos a la variable espectral. Esto, aunado a que no hacemos ninguna operación matricial en el cálculo, permite obtener una gran cantidad de resultados con un costo computacional relativamente bajo, una de las grandes ventajas de nuestro método [4].

3. Resultados

En esta sección ilustraremos la teoría elaborada en la sección previa. Dado que la geometría del sistema queda caracterizada por la función característica $B(\vec{r})$ y que esta toma los valores 1 y 0 únicamente, podemos iniciar nuestro cálculo con la imagen digitalizada de su celda unitaria (ver fig. 9). Esta ima-

gen puede ser escalada, rotada y deformada mediante programas estándares de procesamiento gráfico, con lo cual podemos visualizar la geometría e inmediatamente obtener sus propiedades ópticas. Este aspecto facilita el diseño de nuevos materiales artificiales con propiedades ópticas específicas.

Para ilustrar el poder de nuestro formalismo, en la fig. 10 ilustramos un sistema formado por una red cuadrada de cilindros dieléctricos inmersos en una matriz de oro. Escogimos $\epsilon_B = 4$ como respuesta de los cilindros, mientras que la respuesta del Au $\epsilon_A = \epsilon_{Au}$ fue tomada para distintas frecuencias de resultados experimentales. Elegimos una razón entre ejes principales $r = 1.8$ y una fracción de llenado de $f = V_B/(V_A + V_B) = 0.45$, donde V_A y V_B son los volúmenes ocupados por cada una de las componentes. Consideramos que al meta-material como un sistema semi-infinito cuya superficie es normal a los ejes del cilindro. Dimos valores diversos al ángulo θ formado entre el eje mayor de las elipses y uno de los ejes cristalinos (el eje vertical en la figura) y calculamos el tensor dieléctrico siguiendo los pasos descritos en la sección anterior. Diagonalizamos dicho tensor para identificar las direcciones principales, que denotamos por x y y y los valores principales de la respuesta dieléctrica, y empleamos dichos valores para calcular la reflectividad del sistema cuando es iluminado por luz propagándose a lo largo de los cilindros y polarizada a lo largo de las direcciones principales.

Notamos que para luz con polarización y, la dirección principal más cercana a la dirección vertical en la figura, la reflectancia varía suavemente a lo largo de una curva con la misma forma y que no difiere significativamente de la del Au, aunque para frecuencias entre 1 y 2eV el meta-material *brilla más que el oro*, i.e., su reflectancia es mayor. La reflectancia tiene un mínimo un poco antes que el mínimo del oro cerca de 2.5eV.

En contraste, para polarización x, la dirección principal más cercana a la horizontal en la figura, los resultados son dramáticamente distintos. La reflectancia es alta a frecuencias bajas, como podríamos esperar de cualquier conductor, pero tiene un mínimo extremadamente profundo, donde se comporta como un absorbedor casi perfecto. Variando el ángulo θ entre 9° y 22°, la frecuencia de dicho mínimo puede sintonizarse entre 0.7 y 1.5eV Para ángulos aún mayores, el mínimo queda fijo en 1.5eV y su profundidad disminuye gradualmente. Hay un segundo mínimo fijo alrededor de los 1.6eV cuya profundidad depende de θ. En esta figura no exploramos ángulos menores a 9°, pues las elipses se translaparían entre sí, interrumpiendo los canales conductores en la dirección horizontal y cambiando totalmente el carácter del sistema.

La fig. 11 muestra resultados para un sistema similar, en el que la matriz de Au es reemplazada por una matriz de Ag. Si bien los resultados permanecen cualitativamente similares, el mínimo en la reflectancia se vuelve aún más profundo y se puede sintonizar sobre un rango de frecuencias aún más amplio, desde 1 hasta 2eV.

La fig. 12 muestra resultados para un sistema similar al anterior, pero ahora para una red rectangular uno de cuyos lados mide el doble que el otro. En este caso persisten los resultados cualitativos. El sistema es un reflector casi perfecto para luz con cierta polarización mientras que es un buen absorbedor para la

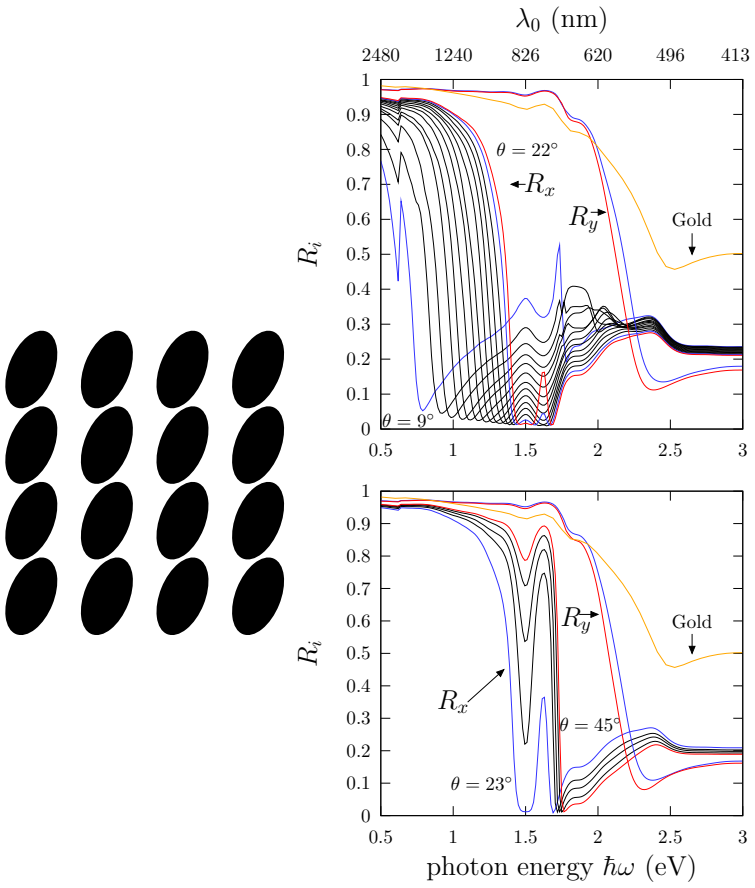


Figura 10: Meta-material formado por cilindros dieléctricos con respuesta $\epsilon_B = 4$ de sección elíptica con razón entre semiejes $r = 1.8$ inmersos en una red cuadrada dentro de una matriz de oro con una fracción de llenado $f = 0.45$ (izquierda). Reflectividad del sistema para luz que incide sobre el material a lo largo del eje de los cilindros con polarización a lo largo de las direcciones principales para diversos ángulos θ entre el eje mayor y un lado de la celda unitaria para θ entre 9° y 22° (derecha arriba) y para θ entre 23° y 45° (derecha abajo). Como referencia, se muestra la reflectancia del oro sólido.

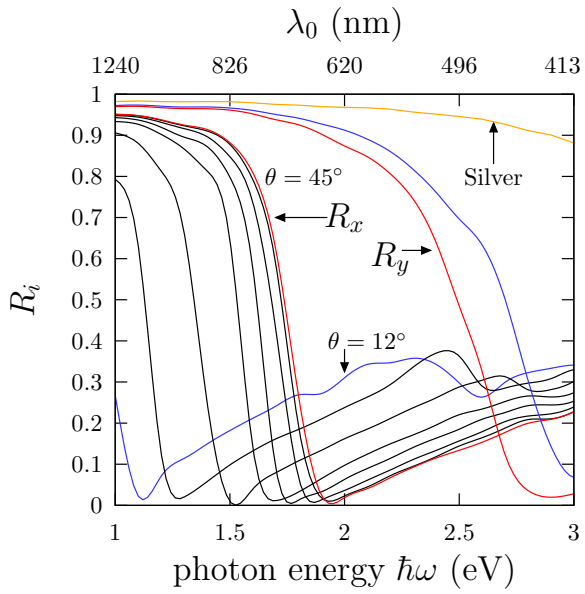


Figura 11: Reflectancia para un sistema similar al de la figura 10, pero reemplazando la matriz de Au por una matriz de Ag.

polarización perpendicular, con un mínimo profundo que puede entonarse en un rango todavía mayor, desde 1 hasta 2.5eV, para el cual el sistema ¡se comporta como un polarizador ideal bajo reflexión!

La absortancia entonable casi perfecta que muestran los sistemas semi-infinitos anteriores están íntimamente ligados con la transmitancia extraordinaria que muestran algunas películas delgadas con arreglos de agujeros. Sin embargo, en nuestra teoría no aparecen explícitamente los plasmones de superficie que suelen invocarse para explicar dicha transparencia. Sin embargo, nuestros cálculos sugieren una explicación cualitativa alternativa. Las figuras anteriores muestran que la absortancia casi-perfecta es aparente para polarizaciones a lo largo de las cuales los caminos conductores se hallan casi estrangulados. Por ejemplo, en la figura 12 la absortancia es grande para polarización cercana a la dirección horizontal, mientras que la reflectancia es alta a lo largo de la dirección vertical. Una explicación cualitativa simple de este fenómeno es la siguiente. Si existen trayectorias conductoras abiertas que vayan de un extremo al otro del sistema, el sistema es un conductor, sin importar qué tan angostos sean los pasos donde éstas se hallen casi estranguladas. La función dieléctrica de un conductor a baja frecuencia es negativa y tiende a $-\infty$ conforme la frecuencia tiende a cero, aunque los caminos conductores sean pobres. Por otro lado, a frecuencia alta, la respuesta dieléctrica está dominada por procesos de polarización, los cuales se ven poco afectados por la existencia de trayectorias conductoras, sobre todo cuando éstas se hallan a punto de estrangularse. Por lo tanto, a frecuencia alta

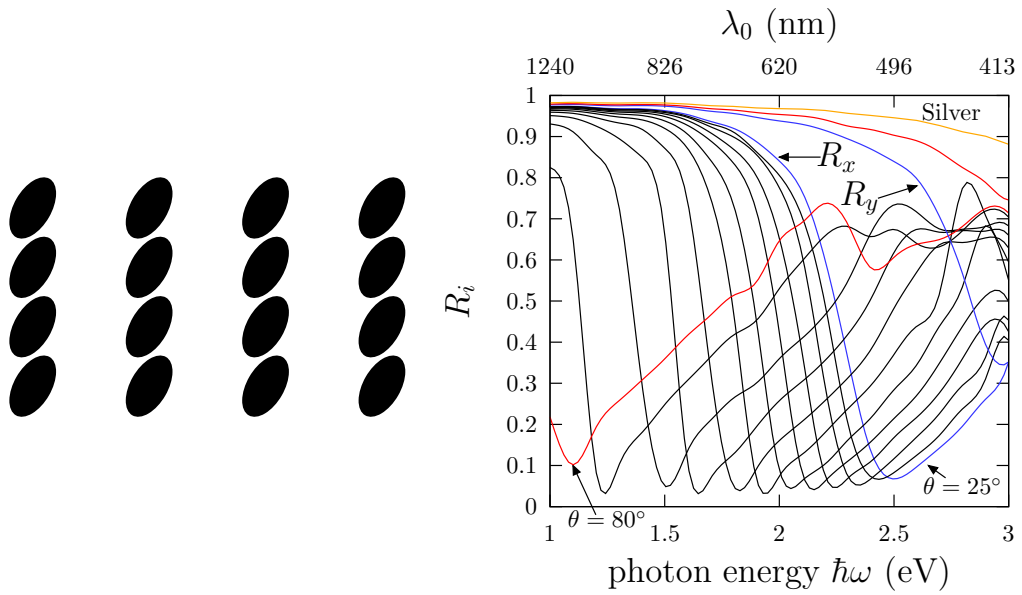


Figura 12: Reflectancia para un sistema similar al de la figura 11, pero en una red rectangular con parámetro de red en una dirección del doble que en la dirección ortogonal y para ángulos que varían entre $\theta = 25°$ y $80°$.

la respuesta se parece a la de un dieléctrico, con valores positivos y con una serie de picos correspondientes a oscilaciones resonantes en las cargas inducidas sobre la superficie de los dieléctricos. Este comportamiento está ilustrado en la figura 13 que muestra la función dieléctrica macroscópica calculada mediante nuestro formalismo para un sistema de prismas rectangulares aislantes inmersos en una matriz de Au para distintos anchos de los canales conductores. Notamos que mientras más largos son los rectángulos dieléctricos (mayor ξ_x) más angostos son los canales conductores y la parte real de ϵ_M alcanza valores positivos más grandes y a frecuencias más bajas. A alguna frecuencia intermedia entre el comportamiento conductor de baja frecuencia y las resonancias dieléctricas de alta frecuencia, la parte real de ϵ_M debe tomar el valor $\epsilon'_M = 1$. Si esto sucede antes de que se haya desarrollado una resonancia, de forma que la parte imaginaria sea pequeña, tendremos un empatamiento de impedancias casi perfecto con la impedancia del vacío, en cuyo caso la reflectancia se abate, casi toda la energía electromagnética se absorbería en un sistema semi-infinito o se transmitiría en una película delgada.

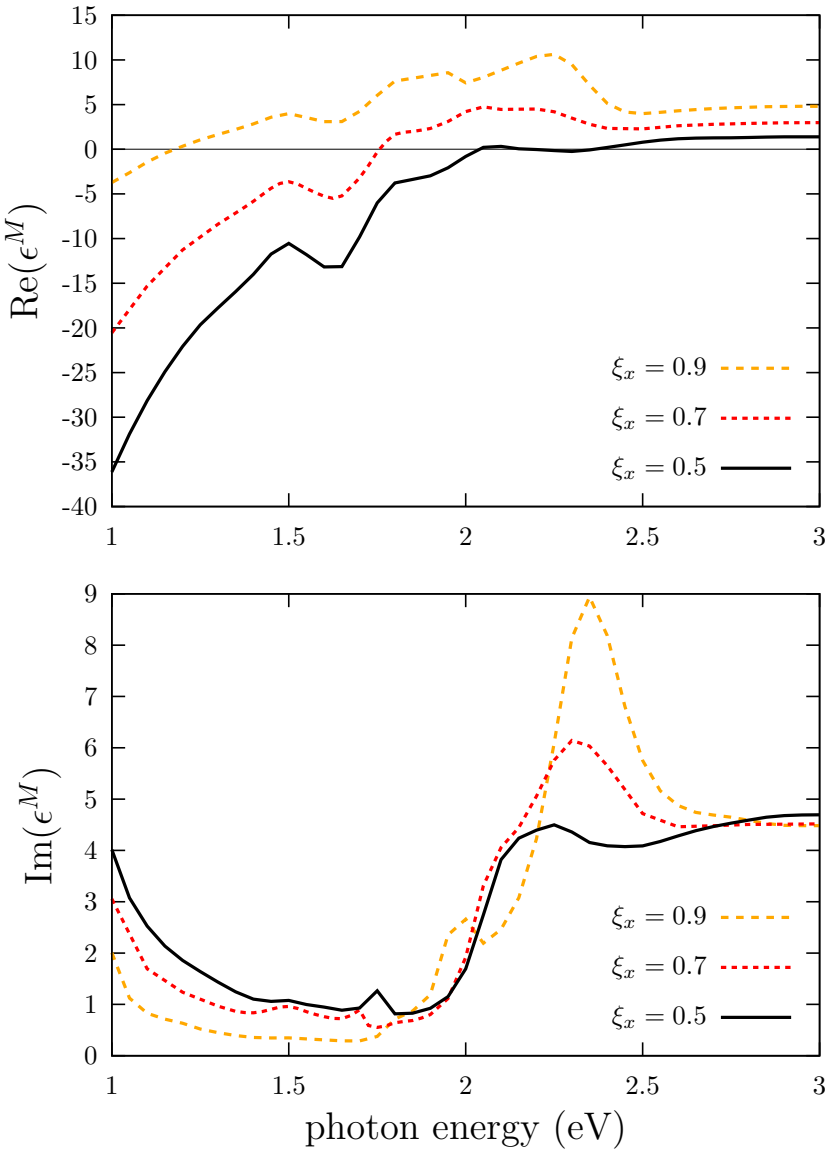


Figura 13: Parte real e imaginaria de la respuesta dieléctrica macroscópica a lo largo de la dirección y para una red cuadrada de prismas dieléctricos rectangulares cuya longitud a lo largo de x es una fracción ξ_x de la celda unitaria.

4. Retardamiento

4.1. Generalización de la teoría

En esta sección generalizaremos nuestra teoría al caso retardado, en el cual no podemos despreciar el tamaño de las inclusiones ni el de la celda unitaria al compararlos con la longitud de onda libre del sistema ni la longitud de onda asociada a la onda de Bloch que se propaga en su seno. Para ello, haremos un desarrollo análogo al que nos condujo a la ec. (7). A partir de las ecuaciones de Maxwell podemos construir una ecuación de onda *microscópica* con fuentes, la cual escribimos como

$$\vec{E} = \frac{4\pi}{i\omega}\hat{\mathcal{W}}^{-1}\vec{J}_{ext} \tag{35}$$

donde $\vec{J}_{ext}$ son las corrientes eléctricas externas y hemos introducido al operador de onda

$$\hat{\mathcal{W}} = \hat{\epsilon} + \frac{c^2}{\omega^2}\hat{P}^T\nabla^2. \tag{36}$$

Suponiendo que las corrientes externas no tienen fluctuaciones, lo cual es indispensable para dar sentido a un planteamiento macroscópico, podemos promediar la ec. (35) para obtener

$$\vec{E}_p = \frac{4\pi}{i\omega}\hat{\mathcal{W}}_{pp}^{-1}\vec{J}_{ext}, \tag{37}$$

lo cual nos lleva de inmediato al *operador de onda macroscópico*

$$\hat{\mathcal{W}}_M^{-1} = \hat{\mathcal{W}}_{pp}^{-1}. \tag{38}$$

Por analogía con el operador de onda microscópico, escribimos

$$\hat{\mathcal{W}}_M = \hat{\epsilon}_M + \frac{c^2}{\omega^2}\hat{P}^T\nabla^2, \tag{39}$$

de donde podemos identificar la respuesta dieléctrica macroscópica del sistema.

El resultado anterior puede generalizarse aún más a todo tipo de funciones respuesta: La respuesta macroscópica a una excitación *externa* es el simple promedio de la correspondiente respuesta microscópica. En el caso electromagnético no retardado longitudinal, la excitación externa es el campo eléctrico externo y la respuesta correspondiente es la inversa de la función dieléctrica longitudinal. En el caso electromagnético genérico, la fuente externa la constituye la corriente eléctrica externa, y la función respuesta es la inversa del operador de onda. Lo único que requerimos para generalizar nuestro resultado a otro tipo de campos es identificar la fuente externa y la función respuesta correspondiente. Así, podemos hallar expresiones para la conductividad térmica macroscópica, para el tensor elástico macroscópico, etc.

4.2. Recursión de Haydock con retardamiento

Para obtener una expansión análoga a la eficiente expansión de Haydock aparecida en el caso no retardado, empezamos por escribir el operador de onda

microscópico en un sistema de dos componentes como

$$\hat{\mathcal{W}} = \frac{\epsilon_A}{u}(u\hat{\xi} - B), \tag{40}$$

donde hemos empleado las mismas definiciones que en la sección 2.3 e introdujimos el operador

$$\hat{\xi} = \hat{1} - \frac{1}{q_A^2}\mathcal{P}^T\nabla^2 \tag{41}$$

y el número de onda libre q_A en el medio A, $q_A = \sqrt{\epsilon_A}\omega/c$. Entonces, el operador de onda inverso puede escribirse como

$$\hat{\mathcal{W}}^{-1} = \frac{u}{\epsilon_a}\hat{g}(u - \hat{B}\hat{g})^{-1}, \tag{42}$$

con $\hat{g} = \hat{\xi}^{-1}$. Identificamos a $\hat{\mathcal{H}} = \hat{B}\hat{g}$ como a nuestro Hamiltoniano, el cual *es* hermitiano si definimos los productos escalares interpretando a $\hat{g}$ como si fuese un *tensor métrico*, i.e. si interpretamos a $\langle\psi|\hat{g}|\phi\rangle$ y no a $\langle\psi|\phi\rangle$ como el producto escalar de los estados $|\phi\rangle$ y $|\psi\rangle$.

Consideramos ahora un cristal artificial, como en la sección 2.2, y construimos el estado macroscópico $|0\rangle$ como una onda plana con $\vec{K} = 0$. Buscamos

$$\hat{\mathcal{W}}_M^{-1} = \frac{u}{\epsilon_a}N^2 g_0(u - B\hat{g})^{-1}_{00}, \tag{43}$$

donde usamos el hecho de que la métrica es diagonal en la representación de ondas planas. El término N^2 en esta ecuación aparece debido al cambio en la normalización de los estados, i.e., a que nuestro estado inicial no es una simple onda plana $\hat{e}\exp(i\vec{k}\cdot\vec{r})$ como en la sección 2.4 y como requerimos para definir *promedio*, con $\hat{e}$ un vector de polarización unitario, sino a $\hat{e}\exp(i\vec{k}\cdot\vec{r})/N$ con N un factor requerido para que el estado $|0\rangle$ esté normalizado respecto a la métrica $\hat{g}$, i.e., $\langle 0|\hat{g}|0\rangle = g_0 = \pm 1$. En analogía a la sección 2.4, aplicamos repetidamente el *hamiltoniano* para construir una base $|\tilde{n}\rangle \equiv \hat{\mathcal{H}}|n-1\rangle = b_n|n\rangle + a_{n-1}|n-1\rangle + g_{n-1}g_{n-2}b_{n-1}|n-2\rangle$ cuyos coeficientes hallamos de orto-normalizar los estados $\langle n|\hat{g}|m\rangle = g_n\delta_{nm}$, donde introdujimos los signos $g_n = \pm 1$ pues la métrica ya no es definida positiva. Obtenemos

$$a_{n-1}g_{n-1} = \langle n-1|\hat{g}|\tilde{n}\rangle \tag{44}$$

y

$$b_n^2 g_n = \langle\tilde{n}|\hat{g}|\tilde{n}\rangle - a_{n-1}^2 g_{n-1} - b_{n-1}g_{n-2} \tag{45}$$

y el signo g_n se sigue de exigir $b_n^2 > 0$. En esta base, el Hamiltoniano es tridiagonal con elementos diagonales a_n, sub diagonales b_n y supra-diagonales $g_n g_{n-1} b_n$. El elemento 00 de la función de Green correspondiente se puede expresar en términos de una fracción continuada como en la ec. (34), simplemente reemplazando b_n^2 por $g_n g_{n-1} b_n^2$.

Siguiendo los pasos esbozados arriba se puede encontrar el operador de onda macroscópico y finalmente el tensor dieléctrico macroscópico. Es importante

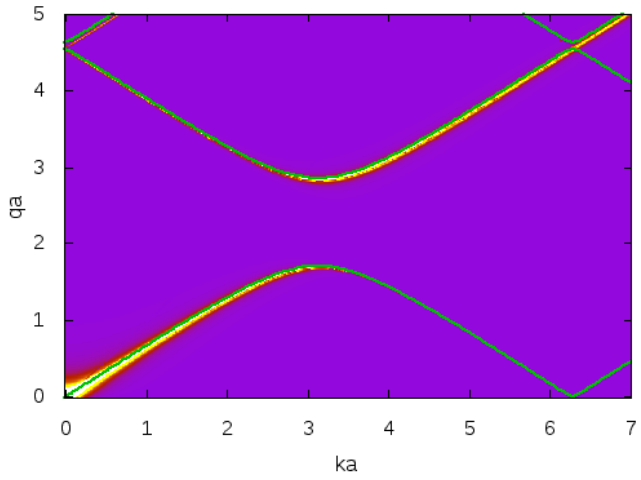


Figura 14: Relación de dispersión q vs $\vec{k}$ para la propagación de una onda transversal a lo largo del eje de una super-red dieléctrica. Se muestra (línea verde) la relación de dispersión exacta, obtenida analíticamente.

hacer notar que ésta es una función respuesta *no local*, la cual depende por separado del vector de onda macroscópico o vector de Bloch $\vec{k}$, cuyo tamaño no hemos despreciado, y del número de onda libre $q = \omega/c$ el cual corresponde a la frecuencia ω. En principio, este cálculo admite vectores y números de onda k y q que pueden ser grandes comparados con el inverso del parámetro de red d.

5. Resultados con retardamiento

Como resultados preliminares del cálculo retardado, y para validar el mismo, calculamos la respuesta dieléctrica no local $\epsilon_M(\omega, \vec{k})$ para un sistema simple unidimensional periódico con periodo $d = \Lambda$ formado por dos películas dieléctricas planas de ancho $d_A = 0.7\Lambda$ y $d_B = 0.3\Lambda$ y con respuestas dieléctricas $\epsilon_A = 1$ y $\epsilon_B = 5$ respectivamente. Para ilustrar los resultados, en la figura 14 mostramos la relación de dispersión $q \equiv \omega/c$ vs k para la propagación de una onda transversal a lo largo del eje de la super-red, la cual obtenemos a partir de $k^2 = \epsilon_M^T(\omega, k)q^2$. En esta figura establecimos el color de cada pixel de acuerdo al inverso de su distancia con la relación de dispersión, i.e., en función de $1/(k^2 - \epsilon_M^T(\omega, k)\omega^2/c^2)$. El caso de una super-red puede resolverse analíticamen-

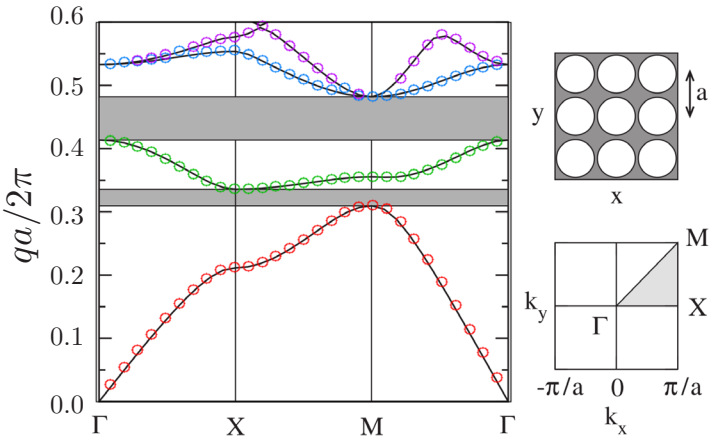


Figura 15: Relación de dispersión de modos TM en una red cuadrada de cilindros huecos $\epsilon_B = 1$ inmersos en una matriz dieléctrica $\epsilon_B = 12$. Se presentan los resultados de nuestro cálculo empleando la respuesta macroscópica y se comparan con los de la referencia [6].

te [5] en términos de la matriz de transferencia M que conecta las proyecciones de los campos $\vec{E}$ y $\vec{H}$ en una interfaz con sus valores en la siguiente interfase equivalente a una distancia d, $\cos kd = \mathrm{Tr}M/2$. La figura muestra que nuestros resultados son consistentes con dicha relación de dispersión. Así, hemos confirmado que nuestro método conduce a la relación de dispersión correcta aún cuando la frecuencia y el número de onda son grandes en la escala determinada por el parámetro de red. En particular, queremos enfatizar que obtuvimos la brecha fotónica en la frontera de la primera zona de Brillouin ($k\Lambda = \pi$). Más aún, nuestros resultados conservan su validez aún más allá de la primera zona de Brillouin. Debemos enfatizar que hubiese sido imposible obtener estos resultados si no hubiésemos tomado en cuenta la dispersión espacial de ϵ_M.

Habiendo establecido que nuestro resultado coincide con el resultado exacto para el caso unidimensional, lo aplicamos a un sistema bi-dimensional para el cual no se conoce una solución analítica. La figura 15 muestra la relación de dispersión para propagación de ondas TM propagándose a lo largo de una red cuadrada de cilindros dieléctricos conforme el vector de onda recorre una trayectoria en la primera zona de Brillouin y su frontera. Los cilindros son huecos ($\epsilon_A = 1$) y están rodeados de un medio dieléctrico (con $\epsilon_B = 12$). El radio de los huecos es de $r = 0.45a$, con a la periodicidad de la red cuadrada. Para cada vector de onda $\vec{k}$ realizamos una búsqueda numérica de aquellas frecuencias ω que satisficieran la relación de dispersión. Comparamos nuestros resultados, indicados con una serie de círculos, con resultados reportados en la literatura para el mismo sistema [6]. Observamos que en general, hay un acuerdo excelente.

6. Conclusiones

En este trabajo desarrollamos un formalismo que nos permite el cálculo eficiente del tensor dieléctrico macroscópico de sistemas periódicos de dos componentes con geometría y composición arbitraria. Hemos aplicado dicho formalismo a sistemas en 1, 2 y 3 dimensiones hechos de todo tipo de redes (cuadradas, rectangulares, cúbicas, orto-rómbicas...) de inclusiones de todo tipo de formas (prismas, cilindros, esferas, cilindros elípticos traslapados...). Nuestro formalismo es aplicable no sólo a sistemas dieléctricos transparentes, sino también a sistemas metálicos, dispersivos y disipativos. En dos dimensiones la representación en el espacio de Fourier de la respuesta microscópica puede requerir del orden de $10^4 \times 10^4$ elementos. Nuestro método de solución basado en una recursión de Haydock no requiere de inversiones ni multiplicaciones matriciales [7], con lo cual el cálculo se acelera más de cinco órdenes de magnitud. La ganancia es aún mayor en el caso tridimensional. Además, nuestro método de solución codifica las propiedades físicas de los materiales que conforman al meta-material en una *variable espectral* cuyo valor puede simplemente substituirse en una fórmula en la cual aparece un conjunto de *coeficientes de Haydock* que dependen exclusivamente de la geometría. Esto permite extender el cálculo a otros materiales y calcular espectros completos sin tener que repetir la parte más cara del cálculo, que es la referente a la geometría. El punto de inicio de nuestro cálculo es un dibujo o una fotografía digitalizada que puede manipularse con herramientas de procesamiento de imágenes para obtener resultados de familias de sistemas con parámetros que pueden sintonizarse para obtener comportamientos específicos. Así, nuestro formalismo permite el *diseño* de meta-materiales con propiedades de interés por sus posibles aplicaciones quizás exóticas, no existentes en materiales naturales.

Como una aplicación, estudiamos la respuesta de un arreglo de cilindros dieléctricos de sección transversal elíptica inmersos en una matriz metálica. Hallamos materiales con una anisotropía extrema consistente en una absortancia muy alta para cierta polarización de la luz incidente y una reflectancia muy alta para la polarización perpendicular. Las frecuencias en las que se presenta este comportamiento se pueden entonar sobre una región extremadamente amplia con solo cambiar la orientación de los ejes principales de las inclusiones. Asimismo, hallamos una explicación de la absortancia extraordinaria, relacionada con la transmitancia extraordinaria observada en películas delgadas metálicas con arreglos de perforaciones, y predijimos que estas son un fenómeno genérico presente siempre que los caminos conductores del sistema se hallen cerca del estrangulamiento.

Generalizamos nuestros resultados para tomar en cuenta el retardamiento en el caso de redes cuyo parámetro fuese comparable con la longitud de onda, y mostramos cómo nuestro formalismo, a pesar de su carácter macroscópico, conduce a la rica estructura de bandas y brechas fotónicas. Comparando con cálculos previos confirmamos la validez de nuestro formalismo aún para vectores de onda en la frontera de la zona de Brillouin y más allá. Sin embargo, para poder aplicarse a frecuencias y vectores de onda grandes, es indispensable tomar

en cuenta que la respuesta macroscópica tiene un carácter no local, originado en la interacción entre partes alejadas del sistema mediante las fluctuaciones del campo electromagnético. En nuestro formalismo macroscópico eliminamos la consideración explícita de estas fluctuaciones, pero no así sus efectos, los cuales conducen a una fuerte dispersión espacial.

En resumen, hemos desarrollado un esquema que permite el cálculo de la respuesta macroscópica de metamateriales periódicos arbitrarios cuya eficiencia computacional permite el diseño de propiedades ópticas exóticas y que captura toda la riqueza de la propagación de ondas electromagnéticas, incluyendo el esquema de bandas y brechas fotónicas.

Referencias

[1] T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio y P. A. Wolff. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**(667), 667-669 (1998).

[2] J. B. Pendry, L. Martin-Moreno y F. J. Garcia-Vidal. Mimicking Surface Plasmons with Structured Surfaces. *Science* **305**(5685), 847-848 (2004).

[3] R Haydock, V Heine y M J Kelly. Electronic structure based on the local atomic environment for tight-binding bands. *Journal of Physics C: Solid State Physics*, **5**(20), 2845 (1972).

[4] W. L. Mochán, Guillermo P. Ortiz y Bernardo S. Mendoza. Efficient homogenization procedure for the calculation of optical properties of 3D nanostructured composites. *Optics Express* **18**(21), 22119-22127 (2010).

[5] Pochi Yeh. Optical waves in layered media. (*Wiley*, 2005).

[6] K. Busch, G. von Freynmann, S. Linden, S.F. Mingaleev, L. Tkeshelashvili y M. Wegener. Periodic nanostructures for photonics. *Physics Reports* **444** 101-202 (2007).

[7] Guillermo P. Ortiz, Brenda E. Martínez-Zérega, Bernardo S. Mendoza y W. Luis Mochán. Effective optical response of metamaterials. *Phys. Rev. B* **79**(24) 245132 (2009).

Fenomenología de Cuerdas

Saúl Ramos-Sánchez

Instituto de Física, Universidad Nacional Autónoma de México,
Apdo. Postal 20-364, México 01000, México

ramos@fisica.unam.mx

Abstract

En estas notas, exploraremos el intrigante mundo de las partículas elementales y de la teoría de cuerdas, concentrándonos en los aspectos de ésta última que podrían permitir establecer una conexión con el modelo estándar de partículas elementales.

1 ¿Hacia dónde vamos? La física moderna

En la física moderna, indudablemente el modelo cosmológico (o modelo ΛCDM) y el modelo estándar de partículas elementales (SM, por sus siglas en inglés) son parte de la herencia más valiosa de la física del siglo XX. Mientras que el primero describe la historia del universo desde la gran explosión hasta nuestra época mediante la aplicación de la relatividad general, el SM describe todas las partículas elementales conocidas en base a la teoría cuántica de campos (QFT, por sus siglas en inglés).

En estas notas, nos concentraremos en el estudio de los avances de la teoría de cuerdas en su camino a describir la física de partículas elementales. Con esta finalidad, esta sección está dedicada al estudio de QFT y del SM, enfatizando aquellos conceptos que nos serán útiles más tarde. En la siguiente sección, describimos los conceptos y las herramientas más importantes de QFT. (Se recomienda [1] como texto introductorio a la teoría de campos.) Posteriormente, dedicamos la sección 1.2 a una descripción de los logros y de los defectos del SM con la finalidad de colocar en contexto el origen de la teoría de cuerdas y su posible alcance.

1.1 Campos y partículas elementales

La QFT está fundamentada en dos conceptos revolucionarios que nacieron entre 1850 y 1950:

- **Campo.** Técnicamente, es una cantidad física asociada a todos los puntos del espacio-tiempo.
- **Simetría.** Es una transformación (de coordenadas, de campos, etc.) que no altera la física de un sistema.

1.1.1 Campos

Entre los campos que nos son más familiares, están los campos magnético $\boldsymbol{B}$ y eléctrico $\boldsymbol{E}$, cuyas líneas de campo se hacen visibles en experimentos sencillos con limadura de hierro. Una observación crucial es que estos campos pueden obtenerse como variaciones de los llamados *potencial escalar eléctrico* ϕ y *potencial vectorial magnético* $\boldsymbol{A}$:

$$\begin{aligned} \boldsymbol{B} &= \nabla \times \boldsymbol{A}\,, \\ \boldsymbol{E} &= -\nabla\phi - \partial_t \boldsymbol{A}\,, \end{aligned} \tag{1}$$

en donde ∂_t denota la derivada parcial con respecto al tiempo. Los potenciales ϕ y $\boldsymbol{A}$ admiten distintos valores en cada punto del espacio tiempo x^μ, en donde el índice de Lorentz corre como $\mu = 0, 1, 2, 3$ con $x^0 = t$. En esta notación (covariante) ambos potenciales se combinan elegantemente en el (mal) llamado *potencial* o *campo electromagnético* $A_\mu(x) = (\phi(x), -\boldsymbol{A}(x))$[1]. El campo electromagnético A_μ contiene toda la información sobre los campos eléctrico y magnético, de tal manera que las ecs. (1) pueden reescribirse como un tensor antisimétrico

$$F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu = \begin{pmatrix} 0 & E_{x^1} & E_{x^2} & E_{x^3} \\ -E_{x^1} & 0 & -B_{x^3} & B_{x^2} \\ -E_{x^2} & B_{x^3} & 0 & -B_{x^1} \\ -E_{x^3} & -B_{x^2} & B_{x^1} & 0 \end{pmatrix}, \tag{2}$$

en donde $\partial_\mu = (\partial_0, \nabla)$. Es de esta manera como J. Maxwell hace más de un siglo comprendió que dos de las *fuerzas fundamentales* de la naturaleza conocidas por sus contemporáneos (la electricidad y el magnetismo) son en realidad distintas manifestaciones de una sola fuerza fundamental: el electromagnetismo. Más interesante aún: Maxwell observó que las leyes de Gauss, Faraday y Ampère en el vacío se podían escribir como la ecuación de onda

$$\partial_\rho \partial^\rho F_{\mu\nu} = 0 \quad \Longleftrightarrow \quad \Box^2 F_{\mu\nu} = (\partial_0^2 - \nabla^2) F_{\mu\nu} = 0\,, \tag{3}$$

con $\partial^\mu = (\partial_0, -\nabla) \neq \partial_\mu$. Y esta observación es la segunda más grande contribución de Maxwell a la física. En ese tiempo, la luz ya había sido descrita como una onda. De un golpe, Maxwell se percató de que el campo electromagnético $A_\mu(x)$ con su ecuación de movimiento (3) describe perfectamente a la luz como una típica onda, con todas las propiedades de éstas.

Pero en la época de Maxwell no se conocía la mecánica cuántica. Con el triunfo de ésta, un nuevo reto se presentó. La ecuación de onda (3) a simple vista no parece describir los cuantos de energía que terminaron llamándose fotones. De hecho, mientras que la teoría clásica sostiene que la energía de la luz depende de su intensidad, las resoluciones a los problemas de la *catástrofe ultravioleta* y del *efecto fotoeléctrico* [2] –que asumen que la luz se propaga en pequeñas cantidades uniformes de energía– demostraron que esta energía depende, más bien, de su frecuencia $E = h\,\nu$. La solución a este conflicto se encuentra en la llamada segunda cuantización. En este formalismo, los distintos modos de Fourier del campo electromagnético $A_\mu(x)$ equivalen a un conjunto de osciladores harmónicos desacoplados, que son cuantizados de manera canónica y cuyos niveles de energía corresponden precisamente a la prediccción de Planck y Einstein, $E_n = nh\nu,\ n \in \mathbb{N}$ [3, 4]. Así, diferentes perturbaciones del campo electromagnético cuántico $A_\mu(x)$ en distintos puntos del espacio-tiempo se traducen en fotones γ (o estados del 'campo fotónico') con distintas propiedades. Por lo tanto, en el formalismo cuántico, un campo es capaz de describir un número infinito de partículas.

[1] En estas notas, usaremos unidades naturales, en las que $\hbar = c = 1$

Esta contribución fue clave para entender por completo el efecto fotoeléctrico, en el que la luz interactúa con los electrones contenidos en los metales. Dirac caracterizó a los electrones (y a sus antipartículas, los positrones [2]) mediante un campo $e(x)$ que satisface la ecuación (de Dirac)

$$(i\gamma^\mu \partial_\mu - m)\, e(x) \;=\; 0\,, \tag{4}$$

en donde γ^μ son las *matrices 4×4 de Dirac* y m es la masa del electrón (ver e.g. [1]). Como en el caso del fotón, múltiples perturbaciones del campo del electrón $e(x)$ conducen a una infinidad de electrones.

Una consecuencia interesante de la descripción de Dirac es que las interacciones de este nuevo campo con $A_\mu(x)$ pueden interpretarse como perturbaciones en los campos que permiten cambios en el número de fotones y electrones pero que conservan la energía y el momento [5]. Este tipo de interacciones son representadas actualmente mediante los llamados *diagramas de Feynman*, como el presentado en la fig. 1. En este diagrama, un electrón es aniquilado por un positrón, permitiendo la emisión de un fotón que carga la energía del par electrón-positrón. El proceso inverso también es permitido por la teoría, i.e. un fotón puede dar origen a un par electrón-positrón[3].

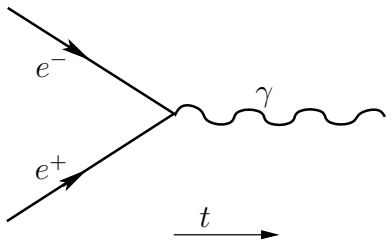


Figure 1: Aniquilación de pares. Interacción de un fotón con partículas cargadas. Se presenta la dirección de evolución en el tiempo t.

Las ecuaciones de movimiento de fotones y electrones se pueden obtener fácilmente aplicando las ecuaciones de campo de Euler-Lagrange

$$\partial_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \psi)} \;=\; \frac{\partial \mathcal{L}}{\partial \psi}\,, \tag{5}$$

en donde ψ representa cualquiera de los campos de un sistema, ψ y su derivada son considerados variables independientes (así como las coordenadas generalizadas q y sus derivadas $\dot{q}$ son independientes en el formalismo clásico), y $\mathcal{L} = \mathcal{L}(\psi, \partial_\mu \psi)$ es llamada *densidad Lagrangiana*, debido a que el Lagrangiano $L = E_{\text{cinética}} - E_{\text{potencial}}$ está dado por

$$L \;=\; \int d^3x\, \mathcal{L} \qquad \Rightarrow \qquad S \;=\; \int dx^0\, L \;=\; \int d^4x \mathcal{L} \tag{6}$$

con S, la acción del sistema. Es fácil mostrar que las densidades Lagrangianas de los campos del fotón y del electrón son

$$\mathcal{L}_e \;=\; \bar{e}\,(i\gamma^\mu \partial_\mu - m)\, e \qquad \& \qquad \mathcal{L}_A \;=\; -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \tag{7}$$

con $F_{\mu\nu}$ definido en ec. (2). La Lagrangiana $\mathcal{L}_{e,A} = \mathcal{L}_e + \mathcal{L}_A$ conduce a las ecuaciones de movimiento (3) y (4) a través de la igualdad de Euler-Lagrange (5) (considerando $\bar{e}$, el positrón, como un campo independiente de e). Sin embargo, en este contexto las interacciones entre fotones y electrones no han sido incorporadas.

[2] Las antipartículas tienen las mismas propiedades que las partículas, salvo por la carga eléctrica, que es opuesta en signo. Por ejemplo, el positrón tiene la misma carga que el electrón pero con signo positivo. Nótese que los fotones no tienen antipartícula.

[3] A los fotones involucrados en estos procesos se les llama virtuales debido a que es preciso que tengan masa no trivial.

Incluir las interacciones electromagnéticas en el formalismo Lagrangiano de la teoría de campos condujo a la teoría de electrodinámica cuántica [6] (QED, por sus siglas en inglés), cuya densidad Lagrangiana es

$$\mathcal{L}_{QED} \;=\; \bar{e}\,(i\gamma^{\mu}D_{\mu} - m)\,e - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}\,, \tag{8}$$

en donde la llamada *derivada covariante* $D_{\mu} = \partial_{\mu} - iq_e A_{\mu}$ (con $q_e > 0$ la carga del positrón) es una extensión de la derivada ∂_{μ} en $\mathcal{L}_e$ que conduce a la densidad Lagrangiana de interacción entre un fotón, un electrón y un positrón asociada al diagrama 1:

$$\mathcal{L}_{int} \;=\; q_e \bar{e}\gamma^{\mu}eA_{\mu}\,. \tag{9}$$

1.1.2 Simetrías

Toda teoría física que actúe en nuestro espacio-tiempo debe ser invariante de Poincaré, es decir, debe permanecer sin modificación cuando se le estudia desde cualquier marco de referencia inercial. El conjunto de matrices 4×4 que describen estas transformaciones de coordenadas forman el llamado *grupo de Poincaré*. Invariancia de Poincaré quiere decir que, si se permite que el grupo de Poincaré actúe de manera *global*, i.e. de manera idéntica en todos los puntos del espacio-tiempo, ninguna de las mediciones realizadas en el sistema original y el resultante será diferente. Es a este tipo de invariancia a lo que se le llama una *simetría global del sistema*.

Toda transformación global (elemento de un grupo de transformaciones globales) puede representarse como

$$U \;=\; e^{i\hat{\alpha}}\,, \tag{10}$$

en donde $\hat{\alpha}$ es, en general, un operador complejo. Es fácil verificar que si U afecta al campo del electrón $e(x)$ en (7) como $e \to e' = U\,e$ y $\bar{e} \to \bar{e}' = U^{\dagger}\,\bar{e}$ con $\hat{\alpha}$ hermitiano, entonces $\mathcal{L}_e$ no cambia. De hecho, tampoco $\mathcal{L}_{QED}$ es afectado y, por tanto, la acción de QED satisface $S \to S' = S$, i.e. es invariante bajo estas transformaciones globales.

Sin embargo, en la discusión anterior, hemos dejado de lado una importante propiedad de las ecuaciones de Maxwell. La alteración de los potenciales clásicos de acuerdo a

$$\phi \to \phi - \frac{1}{q_e}\partial_0\alpha \qquad \& \qquad \boldsymbol{A} \to \boldsymbol{A} + \frac{1}{q_e}\nabla\alpha \tag{11}$$

(con $\alpha = \alpha(x)$ una función (real) arbitraria) conduce a exactamente los mismos campos $\boldsymbol{E}$, $\boldsymbol{B}$ y, por lo tanto, a las mismas ecuaciones de Maxwell. En notación covariante, ec. (11) se reescribe como

$$A_{\mu} \to A_{\mu} - \frac{1}{q_e}\partial_{\mu}\alpha(x)\,. \tag{12}$$

Es necesario destacar algunos puntos importantes:

❶ Debido a que las ecuaciones de Maxwell permanecen invariantes bajo (12), la física no cambia y esta transformación representa una simetría. Este tipo de simetrías que se refieren a grados de libertad debidos a redundancias en la teoría se conocen como *simetrías de norma*.

❷ Es posible reescribir ec. (12) en términos de una transformación similar a U en (10):

$$A_{\mu} \to UA_{\mu}U^{-1} - \frac{1}{q_e}\partial_{\mu}(U)U^{-1}\,. \tag{13}$$

En este caso, sin embargo, debemos emplear una *transformación local* $U = U(x) = e^{i\alpha(x)}$. Es muy importante notar la diferencia entre las simetrías *globales* y *locales*. Mientras que las primeras actúan en cada punto del espacio-tiempo de manera idéntica, una transformación local transforma los campos de manera distinta dependiendo del punto del espacio-tiempo en el que éste esté definido.

❸ En la teoría cuántica de QED, $\mathcal{L}_{QED}$ permanece invariante si además de (13) el campo del electrón se transforma como

$$e \to Ue\,. \tag{14}$$

En conjunto, ecs. (13) y (14) describen una simetría de norma en QFT. La densidad Lagrangiana asociada describe entonces una *teoría (local) de norma*. Nótese que, a pesar de que las transformaciones locales actúan diferente en cada punto del espacio-tiempo, la teoría completa es invariante bajo la simetría.

❹ En el lenguaje de teoría de grupos,[4] si $U(x) = e^{i\alpha(x)}$ representa el único generador independiente de transformaciones con $\alpha : \mathbb{R}^4 \mapsto \mathbb{R}$, entonces el grupo de simetría asociado es U(1). En el caso del electromagnetismo, denotaremos el grupo de simetría como $\mathrm{U}(1)_{QED}$. U(1) es un grupo abeliano, es decir, dos elementos U y $\widetilde{U}$ de U(1) conmutan, $[U, \widetilde{U}] = 0$. Sin embargo, es posible construir teorías de norma basadas en grupos no-abelianos (donde las transformaciones no conmutan).

❺ En las teorías de norma, los bosones asociados a los campos que se transforman como (13) se les conoce como *bosones de norma* y tienen espín $s = 1$. Por lo tanto, el fotón es el bosón de norma de la QED. Los bosones de norma pueden interpretarse como los mediadores de las interacciones entre las partículas (fermiónicas, i.e. con espín semientero) que se transforman como en (14). Las interacciones aparecen, como en ec. (8), a través de la definición de la derivada covariante D_μ, la cual contiene a todos los bosones de norma de la simetría en cuestión. En general, las simetrías de norma no-abelianas conducen a múltiples bosones de norma, mientras que e.g. $\mathrm{U}(1)_{QED}$ tiene un sólo bosón de norma: el fotón.

Las simetrías de norma han tenido un éxito rotundo en la descripción de la física de partículas elementales. Como veremos a continuación, el SM se basa en una combinación simetrías de norma abelianas y no-abelianas para describir todas las partículas que componen la materia observada y sus interacciones.

1.2 El modelo estándar de partículas elementales y sus problemas

Nuestro universo no está compuesto sólo de electrones y fotones. Observaciones de partículas altamente energéticas que llegan a nuestro planeta desde regiones muy distantes (llamadas *rayos cósmicos*) permitieron saber que existen otras partículas que sólo se distinguen del electrón en su masa (i.e. tienen la misma carga eléctrica y aparecen en el mismo tipo de fenómenos físicos); a éstas se les llamó *leptones*. Simultáneamente, se descubrió que además existen otras partículas mucho más pesadas llamadas *hadrones* que, de cierta forma, se parecen más a los constituyentes de los núcleos atómicos. Como veremos aquí, mientras que los leptones son partículas fundamentales[5], los hadrones están compuestos de otras partículas más pequeñas.

En la sección anterior, hemos visto que electrones y fotones interactúan y que estas interacciones son descritas por una teoría de norma con grupo de simetría $\mathrm{U}(1)_{QED}$. Sin embargo, no

[4] Se sugiere [7, 8] como útiles revisiónes de teoría de grupos con énfasis en su papel en la física de partículas elementales.

[5] 'Fundamental' en este contexto quiere decir que ningún experimento ha sido capaz de mostrar que es posible descomponer a las partículas fundamentales en pequeños fragmentos más básicos (aunque, como veremos, la teoría de cuerdas indica lo contrario).

sólo los electrones perciben las interacciones electromagnéticas. Existen también neutrinos (con carga neutra), protones (con carga positiva), neutrones (sin carga), etc. que también interactúan electromagnéticamente. En principio, nada impide la generalización de $\mathcal{L}_{QED}$ para incluir campos similares a $e(x)$ relacionados con estas partículas adicionales, ya que, si analizamos estas partículas de manera aislada, salvo por su masa y carga, todas ellas son idénticas.

(Des)Afortunadamente, la historia no es tan sencilla. La física nuclear nacida a principios del siglo pasado, tras el descubrimiento de los decaimientos radiactivos, es la 'manzana de la discordia'. Mientras que la teoría de norma $\mathrm{U}(1)_{QED}$ permite entender cómo la luz interactúa con los electrones atómicos, no explica la emisión de un electrón cuando un núcleo de ^{137}Cs se transforma en ^{137}Ba. En este proceso conocido como *decaimiento* β, un neutrón se transforma en un protón dentro del núcleo, provocando la transición nuclear y la emisión de una partícula β (como se le bautizó al electrón emitido en este proceso) muy energética, capaz de destruir otros átomos en su trayectoria (es este un origen del cáncer y del envenenamiento radiactivo). Adicionalmente, las interacciones electromagnéticas no explican la existencia de núcleos estables, ya que el electromagnetismo predice que varias partículas igualmente cargadas (como los protones) se repelen. Estos fenómenos nucleares encontraron explicación cuando se entendió que los protones y neutrones que construyen los núcleos están compuestos a su vez de otras partículas *más elementales* que hoy llamamos *quarks* y que se rigen por reglas adicionales a las del electromagnetismo.

Los ingredientes extra son las fuerzas fundamentales involucradas en las interacciones nucleares. Hoy se sabe que existen dos fuerzas que operan a distancias tan pequeñas como el radio de un núcleo (aproximadamente 10^{-13} cm). La *fuerza fuerte* establece sólidos enlaces entre los quarks que constituyen a los protones y neutrones. La *fuerza débil*, en cambio, provoca transiciones de quarks a leptones (como los electrones), es decir, es responsable del decaimiento radiactivo. El trabajo realizado durante la primera mitad del siglo pasado consistió, pues, en entender estas nuevas fuerzas. La conclusión fue que, al igual que el electromagnetismo, estas fuerzas son descritas por medio de teorías de norma. En particular, la fuerza débil es caracterizada por el grupo no-abeliano $\mathrm{SU}(2)_L$, mientras que las interacciones de la fuerza fuerte encuentran su descripción en el grupo de transformaciones no-abelianas $\mathrm{SU}(3)_c$.

Las transformaciones débiles son descritas con el mismo grupo que describe las transiciones de espín, SU(2). Por lo tanto, es de esperarse que las partículas involucradas en las interacciones débiles tengan una carga débil similar al espín. Esta carga es denominada *isoespín débil*. En analogía con el espín, algunas partículas tienen isoespín $+\frac{1}{2}$ y pueden tener transiciones a partículas con isoespín $-\frac{1}{2}$. Se dice que este tipo de partículas forman un doblete de isoespín débil o que se transforman como una representación **2** del grupo $\mathrm{SU}(2)_L$. Las transiciones entre partículas que forman representaciones de $\mathrm{SU}(2)_L$ son mediadas por 3 bosones de norma llamados bosones débiles $W_\mu^{1,2,3}$. La magnitud de las interacciones débiles está determinada por la 'constante' de acoplamiento g_2, de tal manera que la transformación de un leptón ℓ y un bosón de norma bajo la simetría $\mathrm{SU}(2)_L$ se expresa como

$$\ell \to U\ell\,, \qquad W_\mu^i \to U W_\mu^i U^{-1} - \frac{1}{g_2}\partial_\mu(U)U^{-1}\,, \qquad U(x) \;=\; e^{i\alpha^i(x)t_i}\,, \tag{15}$$

donde $t_i = \frac{1}{2}\sigma_i$ con σ_i las matrices de Pauli. Nótese que, dado que las matrices de Pauli no conmutan, distintas transformaciones U tampoco, razón por la que SU(2) es un grupo no abeliano. La Lagrangiana de interacción en este caso es

$$\mathcal{L}_{int} \;=\; g_2\bar{\ell}\gamma^\mu\ell W_\mu^i t_i\,. \tag{16}$$

Una observación adicional sobre estas interacciones tiene qué ver con la llamada quiralidad. A grandes rasgos, una partícula *izquierda* (*derecha*) con campo ψ_L (ψ_R) es aquella cuya orientación de

giro espinorial se describe con la mano izquierda (derecha), donde el pulgar apunta en la dirección de desplazamiento. En principio, un campo arbitrario tiene ambas componentes. Una teoría es quiral si las componentes izquierdas se transforman de manera diferente a las componentes derechas bajo alguna simetría. El índice del grupo $\mathrm{SU}(2)_L$ tiene su origen en el hecho de que sólo partículas izquierdas son capaces de transformarse. Esto implica que la teoría de norma de las interacciones débiles es una teoría quiral. Como ejemplo, mencionemos las componentes izquierdas de los quarks que forman un neutrón, un quark arriba u (up) y dos quarks abajo d (down), y los quarks de un protón, 2 quarks u y un quark d. Las componentes izquierdas de éstos forman un doblete de $\mathrm{SU}(2)_L$; entonces, transiciones $u_L \longleftrightarrow d_L$, que involucran la emisión de un bosón débil W_μ, son precisamente el origen del decaimiento β: un neutrón $u_L d_L d_L$ sufre una transición a $u_L u_L d_L + W_\mu$. El bosón débil liberado decae eventualmente en el electrón (y un neutrino) observado en este decaimiento radiactivo.

El grupo de transformaciones fuertes o grupo *de color* actúa sobre los componentes fundamentales de todos los hadrones: los quarks. Estos aparecen en tres variedades diferentes cada uno que se han llamado, meramente por convención, *colores* y que usualmente se escogen como rojo, verde y azul. Transformaciones entre estos tres colores ocurren todo el tiempo a través del intercambio de los 8 bosones de norma de las interacciones fuertes llamados *gluones* g^i_μ, $i = 1, \ldots, 8$. En este caso, la magnitud de las interacciones es g_3, por lo que las transformaciones de norma de $\mathrm{SU}(3)_c$ de quarks q y gluones están dadas por

$$q \to Uq\,, \qquad g^i_\mu \to U g^i_\mu U^{-1} - \frac{1}{g_3}\partial_\mu(U)U^{-1}\,, \qquad U(x) \;=\; e^{i\alpha^i(x)\lambda_i}\,, \tag{17}$$

donde λ_i son las matrices de Gell-Mann, que no conmutan. La Lagrangiana de interacción en este caso es

$$\mathcal{L}_{int} \;=\; g_3 \bar{q}\gamma^\mu q g^i_\mu \lambda_i\,. \tag{18}$$

La peculiaridad de la fuerza fuerte es que confina a los quarks a vivir juntos, tan apretados como para formar sólidos hadrones. Es decir, los gluones unen muy estrechamente a los quarks, tal que éstos no pueden separarse. Esto se debe a que g_3 es mayor que la unidad a bajas energías (o distancias grandes – mayores que el radio nuclear), mientras que, a energías muy altas o distancias subnucleares, g_3 es tan pequeña como la carga del electrón, permitiendo así que los quarks sean libres dentro del núcleo[6]. Hemos visto ya dos ejemplos de hadrones: los protones y neutrones. Existen muchos otros hadrones que han sido detectados en los rayos cósmicos y en las colisiones de aceleradores de partículas, tales como los piones y kaones. Los colores de los quarks se combinan, de tal forma que los hadrones no tienen color. Por ejemplo, una combinación válida de quarks en un protón es: $u_L^{\text{rojo}} u_L^{\text{verde}} d_L^{\text{azul}}$.

En el lenguaje del grupo asociado a esta fuerza, los quarks forman tripletes **3** (y antitripletes $\overline{\mathbf{3}}$) de $\mathrm{SU}(3)_c$. Así como las componentes de un **2** de $\mathrm{SU}(2)$ pueden transformase entre ellas (como $s = \frac{1}{2} \to s = -\frac{1}{2}$), distintos colores se transforman entre ellos, de tal forma que un protón no puede contener sólo la combinación $u_L^{\text{rojo}} u_L^{\text{verde}} d_L^{\text{azul}}$, sino también $u_L^{\text{azul}} u_L^{\text{verde}} d_L^{\text{rojo}}$ y todas las que conduzcan a un color neutro.

Una vez adquirida cierta familiaridad con las fuerzas fundamentales, es preciso hablar de los componentes básicos de la materia observada. Como anticipamos en el lado de los leptones, el electrón (campo $e(x)$) tiene un 'hermano' sin carga eléctrica llamado neutrino del electrón ($\nu^e(x)$). Existen adicionalmente dos copias de esta pareja: el muón (campo $\mu(x)$) con su neutrino ($\nu^\mu(x)$), y el tauón (campo $\tau(x)$) con su neutrino ($\nu^\tau(x)$). Los neutrinos aparecen generalmente

[6] A esta propiedad se le llama *libertad asintótica*.

		Generaciones					
		Primera	Segunda	Tercera	$SU(3)_c$	$SU(2)_L$	$U(1)_Y$
Fermiones $s = \frac{1}{2}$	Leptones	$(\nu_L^e,\, e_L)$	$(\nu_L^\mu,\, \mu_L)$	$(\nu_L^\tau,\, \tau_L)$	**1**	**2**	$-\frac{1}{2}$
		e_R	μ_R	τ_R	**1**	**1**	1
		ν_R^e	ν_R^μ	ν_R^τ	**1**	**1**	0
	Quarks	$(u_L,\, d_L)$	$(c_L,\, s_L)$	$(t_L,\, b_L)$	**3**	**2**	$\frac{1}{6}$
		u_R	c_R	t_R	$\overline{\mathbf{3}}$	**1**	$-\frac{2}{3}$
		d_R	s_R	b_R	$\overline{\mathbf{3}}$	**1**	$\frac{1}{3}$
Bosones	$s = 1$	B_μ bosón de hipercarga			**1**	**1**	0
		$W_\mu^{1,2,3}$ bosones débiles			**1**	**3**	0
		$g_\mu^{1,\ldots,8}$ gluones			**8**	**1**	0
	$s = 0$	H bosón de Higgs			**1**	**2**	$-\frac{1}{2}$

Table 1: Partículas elementales y sus cargas bajo las simetrías de norma del SM. Los índices $_L$ y $_R$ se refieren respectivamente a la quiralidad izquierda y derecha de las partículas en cuestión. s denota el espín.

simultáneamente con sus compañeros cargados. Por ejemplo, ν^e aparece junto con el electrón en el decaimiento β. Por otra parte, hay también 3 copias de parejas de quarks: los quarks arriba $u(x)$ y abajo $d(x)$ (que constituyen a los protones y neutrones), los quarks 'charm' $c(x)$ y 'strange' $s(x)$, y los quarks 'top' $t(x)$ y 'bottom' $b(x)$. Cada pareja de leptones se asocia a una pareja de quarks para formar una *generación* de quarks y leptones. Por ejemplo, la primera generación está formada por las parejas $e - \nu^e$ y $u - d$; esta generación es la que constituye la materia estable de la naturaleza. Un resumen de todas las partículas elementales, según el SM aparece en la tabla 1. (En esta tabla, no se hace mención de $U(1)_{QED}$, sino de la hipercarga $U(1)_Y$ que, como veremos a continuación, es de donde surge el electromagnetismo.) Hasta hoy, los experimentos muestran que estas partículas son elementales (i.e. que no tienen subestructura) y que no existen partículas adicionales que interactúen con éstas a través de las fuerzas fundamentales conocidas.

Una partícula que merece una mención por separado es el bosón de Higgs. En el SM, la partícula de Higgs es la única partícula elemental con $s = 0$ (las demás tienen $s = \frac{1}{2}, 1$). Esta es una de sus peculiaridades. Por si fuera poco, no ha sido detectada aún, y, más importante aún, en el SM es la raíz de las masas de toda la materia que conocemos, la cual nace mediante el llamado *mecanismo de Higgs.*

1.2.1 Mecanismo de Higgs

No hemos mencionado dos aspectos muy importantes del SM como lo hemos descrito hasta ahora: i) entre las fuerzas fundamentales incluídas en el SM no se encuentra el electromagnetismo, y ii) todas las partículas (fermiones y bosones) carecen de masa. Una consecuencia inmediata es que las interacciones débiles (responsables de la radiactividad) son de alcance infinito, i.e. un electrón emitido por un televisor en casa podría desencadenar una mayor combustión en el sol o en una estrella mucho más lejana. Esto es una contradicción directa a la evidencia experimental que indica que las interacciones débiles sólo ocurren en el núcleo atómico. Por otra parte, no es la carga eléctrica, sino la *hipercarga* q_Y la que aparece en el SM. La hipercarga es una fuerza muy similar al electromagnetismo: es una simetría de norma abeliana $U(1)_Y$ bajo la cual las partículas del SM están cargadas (ver tabla 1) y se acoplan al nuevo bosón de norma B_μ con magnitud g_Y. De hecho,

estamos a punto de ver que el electromagnetismo surge como remanente de las simetrías $\mathrm{SU}(2)_L$ y $\mathrm{U}(1)_Y$ en el estado básico (el *vacío*) de la teoría. Entonces, para completar la descripción de nuestro universo, se requiere de un elemento más.

En el mecanismo de Higgs, el universo sufre una transición de fase que altera su constitución y la dinámica de la materia en él. Esta transición está originada por en la densidad Lagrangiana de la partícula de Higgs:

$$\mathcal{L}_{Higgs} \;=\; (D^\mu H)^\dagger\,(D_\mu H) - V(H)\,, \qquad V(H) \;=\; -\frac{\mu^2}{2}H^\dagger H - \frac{\lambda}{4}(H^\dagger H)^4\,, \tag{19}$$

en donde $V(H)$ es la densidad de energía potencial de Higgs (o simplemente el potencial de Higgs), con μ y $\lambda > 0$ parámetros de la teoría. Además, la derivada covariante D_μ contiene todas los generadores de las simetrías bajo las que el Higgs no se transforma trivialmente (i.e. los generadores de $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$): $D_\mu = \partial_\mu - iq_Y B_\mu - ig_2 t_i W_\mu^i$ con suma sobre $i = 1, 2, 3$ y $t_i = \frac{1}{2}\sigma_i$.

En el SM, originalmente $\mu^2 > 0$. En este escenario, el mínimo de $V(H)$ ocurre en $\langle H\rangle = 0$. A este valor se le llama el valor de expectación o valor esperado de H y corresponde a su valor natural en el vacío del universo. La transición ocurre cuando, dinámicamente, $\mu^2 > 0 \to \mu^2 < 0$. En este proceso el mínimo original se transforma en un máximo y se forma un valle de mínimos (ver fig. 2). Recordando que un sistema física tiene a su nível de mínima energía, tras la transición H adquirá el nuevo valor de expectación $v \equiv |\langle H\rangle| = \sqrt{-\mu^2/\lambda}$. Este cambio minúsculo altera enormemente la teoría.

El nuevo vacío tiene varias repercusiones. Debido al primer término (el término cinético) de (19), aparecen términos de la forma

$$m_W^2 W_\mu^+ W^{\mu-} + m_Z^2 Z_\mu Z^\mu\,,$$

en donde Z_μ y $W_\mu^\pm$ son combinaciones de los bosones de norma originales, W_μ^i, B_μ, y las masas están dadas por $m_W^2 = (vg_2)^2/2$ y $m_Z^2 = \frac{1}{2}v^2(g_Y^2 + g_2^2)$. Estas masas han sido medidas experimentalmente: $m_Z \sim 90$ GeV y $m_W \sim 80$ GeV [7]. Este hecho debe ser subrayado. Como vimos antes, si los bosones de norma de las interacciones débiles carecieran de masa (como los originales de $\mathrm{SU}(2)_L$), las interacciones deberían tener un alcance infinito. En cambio, cuando los bosones son masivos, es decir, tras la transición de fase, las interacciones tienen un alcance bastante reducido debido a la enorme masa de los nuevos bosones $Z_\mu, W_\mu^\pm$. Sin embargo, no sólo los bosones de norma adquieren masa. En el SM, existe un término de interacción entre los fermiones ψ_i y el bosón de Higgs llamado *término de interacción de Yukawa* que, esquemáticamente, tiene la forma

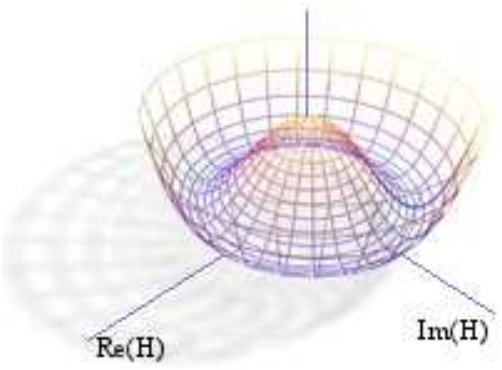


Figure 2: El potencial de Higgs $V(H)$.

$$\mathcal{L}_{Yuk} \;=\; y_i H \overline{\psi}_i \psi_i\,. \tag{20}$$

Cuando H adquiere un valor de expectación $v \neq 0$, la masa de ψ_i de acuerdo a ec. (20) es $m_i = y_i v$. Los valores de y_i y de v han sido determinados experimentalmente a través de la medición de las masas de las partículas del SM. Curiosamente, estos *acoplamientos de Yukawa* y_i aumentan en valor con la generación de quarks y leptones que se considere, i.e. y_i es mucho más pequeño para el electrón que para el muón, y el valor para el muón es más pequeño que para el tauón. Sólo las

[7] En la física de partículas elementales, asumiendo c la famosa ecuación de Einstein establece $E = m$ y, por lo tanto, se prefiere medir las masas en términos de energía.

partículas de la generación más ligera (la primera) son estables, las demás decaen a sus copias más ligeras.

Existe una tercera consecuencia del mecanismo de Higgs que tiene qué ver con las simetrías de la teoría. Dado que el Higgs tiene hipercarga $q_Y = -\frac{1}{2}$ y se transforma como un doblete bajo $\mathrm{SU}(2)_L$, cuando H es reemplazado por su valor de expectación en el nuevo vacío de la teoría, estas simetrías dejan de ser válidas al 100%. Esto quiere decir que en el vacío, de hecho la simetría $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ es "reemplazada" por $\mathrm{U}(1)_{QED}$ en donde la carga electromagnética de cualquier partícula (quark o leptón) puede expresarse como $q_{QED} = q_Y + t_3$, en donde t_3 se refiere al eigenvalor del operador $\frac{1}{2}\sigma_3$ actuando sobre cualquier campo del SM (e.g. en la notación de la tabla 1, ν^e tiene $t_3 = \frac{1}{2}$ mientras que $q_Y = -\frac{1}{2}$; esto conduce a $q_{QED}(\nu^e) = -\frac{1}{2} + \frac{1}{2} = 0$ y $q_{QED}(e) = -\frac{1}{2} - \frac{1}{2} = -1 = -q_e$). Es preciso mencionar que, además de los bosones de norma masivos, la combinación $g_Y W^3_\mu + g_2 B_\mu$ permanece sin masa. Es precisamente esta combinación la que puede identificarse con el campo del fotón A_μ y que conduce, en el vacío, a la simetría de norma válida $\mathrm{U}(1)_{QED}$.

Antes de concluir esta discusión, es útil mencionar que, a pesar de que en el vacío las simetrías originales han desaparecido, éstas siguen teniendo su presencia en la teoría completa (i.e. fuera del vacío). A este tipo especial de rompimiento de simetrías se le llama *rompimiento espontáneo*; en el SM la *simetría electrodébil* $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ ha sido rota espontáneamente a $\mathrm{U}(1)_{QED}$ mediante el mecanismo de Higgs, en el que éste adquiere un valor de expectación.

1.2.2 Los problemas del modelo estándar

El SM es increíblemente exitoso; quizá representa la teoría que tiene las predicciones más precisas. De los 26 parámetros libres del SM (masas de las partículas, valor esperado del Higgs, proporción de mezclas entre quarks y leptones, etc.), sólo la masa del Higgs no ha sido confirmada. Es este precisamente el punto débil del SM: la partícula de Higgs no aparece en los experimentos. De hecho, el gran colisionador de hadrones (LHC, por sus siglas en inglés)[8] ha sido expresamente diseñado y construido para descubrir esta partícula cuya existencia fue predicha hace casi 50 años. Hasta ahora, otros colisionadores han fallado en la búsqueda del Higgs; sin embargo han establecido cotas inferiores y superiores para la masa que el Higgs (con las propiedades indicadas por el SM) puede tener. En resumen, se espera que el Higgs tenga una masa en el intervalo 115 GeV $\lesssim m_H \lesssim$ 145 GeV (ver fig. 3) y que, por lo tanto, el LHC pueda encontrarlo en los próximos años. Sin embargo, cabe la posibilidad de que el Higgs no exista. En ese caso, la pregunta que nacería es ¿cómo se genera la masa de las partículas?

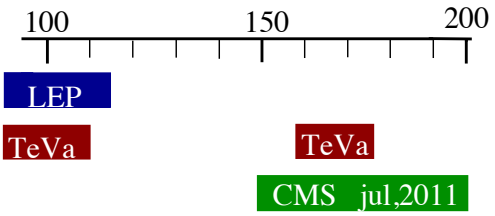


Figure 3: Experimentos en colisionadores de partículas actuales (LEP, TeVatron y CMS del LHC) restringen la masa del Higgs a 115 GeV $\lesssim m_H \lesssim$ 145 GeV. La escala está en GeV. Las regiones coloreadas han sido excluídas por los distintos experimentos.

Otra pregunta que deja sin respuesta el SM es ¿por qué la masa del Higgs, en caso de que éste exista, es tan pequeña? En principio, si el SM es válido en cualquier rango de energía, se presenta una encrucijada parecida a la centenaria catástrofe ultravioleta: la masa del Higgs adquiere tantas contribuciones que crece sin límite. Este problema es llamado *el problema de la jerarquía*. Una forma de visualizar esta jerarquía es, recordando que la escala a la que se cree que las interacciones gravitacionales se vuelven relevantes para un sistema cuántico es la escala energética de Planck $M_{Pl} \sim 10^{19}$ GeV, mientras que la masa del Higgs se espera que sea

[8] El LHC es un colisionador de protones formado por un anillo de 27 km de diámetro y localizado en el laboratorio CERN, en Ginebra, Suiza.

$m_H \sim \mathcal{O}(100)$ GeV. ¿Por qué existe esa enorme diferencia? ¿No hay explicación? Aunque ésta última es ciertamente una posibilidad, la mayoría de las veces semejante respuesta ha sido sustituida por respuestas donadas por teorías tan impactantes como la mecánica cuántica o la relatividad. Entre las propuestas para resolver este conflicto se encuentran la inclusión de una simetría nueva llamada *supersimetría* (que será descrita más tarde) y considerar los efectos de posibles dimensiones adicionales (más allá de las 3 espaciales más la temporal que nos son familiares).

Entre las interrogantes que el SM no explica, existen algunas preguntas vinculadas con los experimentos, sin embargo no las consideraremos aquí, ya que no alteran sustancialmente al SM y podrían estar relacionadas a nuestro desconocimiento de las interacciones fuertes a bajas energías. Otras incógnitas de origen más teórico son: ¿por qué hay tantos (¡26!) parámetros libres en el SM? ¿Por qué hay 3 y no más o menos generaciones de quarks y leptones? ¿Por qué hay diferencias en la masa de las partículas de las distintas generaciones? ¿De dónde surgen las simetrías de norma? ¿Por qué hay algunos parámetros que son prácticamente cero? La máxima pregunta es, empero, ¿cómo se puede incluir la gravedad en el lenguaje empleado para las otras fuerzas? Mientras que 3 de las 4 fuerzas fundamentales han sido descritas cuánticamente (en base a teorías de norma y QFT), la gravedad ha demostrado que no puede ser cuantizada de manera directa. Así, responder esta última pregunta requiere de un replanteamiento serio de la física moderna que, según algunos, pasa por una modificación de la relatividad general o de la mecánica cuántica, o, según otros, se basa en alterar los paradigmas establecidos por el éxito de éstas dos teorías. La teoría de cuerdas está basada en la segunda perspectiva. En particular, sugiere abandonar la idea de que las partículas del SM son elementales, indicando que las partículas no son sino distintas manifestaciones de diminutas cuerdas rotantes. Como veremos más adelante, esta simple idea conduce a una cuantización de la gravedad y, simultáneamente, propociona un origen común para todas las fuerzas fundamentales y las partículas elementales, incluyendo al *gravitón*, el símil gravitacional del fotón en las interacciones electromagnéticas.

2 Teoría de cuerdas y su fenomenología

Esta sección no intenta ser una revisión exhaustiva del enorme progreso realizado en teoría de cuerdas durante las últimas tres décadas. Aquí sólo estudiamos algunas de las propiedades más relevantes de esta teoría, enfocándonos en las herramientas que permiten intentar dar solución a algunas de las preguntas enumeradas en la sección anterior. Para aquellos estudiantes de posgrado interesados en aprender más sobre teoría de cuerdas, recomendamos el trabajo pionero de Green, Schwarz y Witten [9, 10] y el texto de Polchinski [11, 12]. Para estudiantes más jóvenes, es recomendable leer el texto de Balin y Love [13]. A un nivel menos académico, es también gratificante leer [14].

2.1 Introducción a la teoría de cuerdas

2.1.1 La cuerda bosónica

En su concepción clásica, la teoría de cuerdas es una teoría cuántica de campos en la que las partículas puntuales son reemplazadas por diminutos objetos unidimensionales (con longitud proporcional a la *longitud de Planck* $\ell_{str} \propto \ell_{Pl} \sim 10^{-33}$ cm), cuyos modos vibracionales (modos de Fourier) son interpretados como diferentes partículas a las distancias a las que podemos medir[9]. Mientras que la posición de una partícula en D dimensiones puede ser descrita mediante D grados

[9]Para comparar, el núcleo atómico mide aproximadamente 10^{-13} cm y el radio de un electrón es usualmente considerado ligeramente menor, aunque más grande que 10^{-20} cm.

de libertad $X^M(\tau)$, $M = 0, \ldots, D-1$, que dependen sólo del parámetro temporal τ, para describir una cuerda dinámica requerimos añadir una coordenada espacial σ, tal que $X^M(\sigma, \tau)$ traza una curva (la cuerda) cuando σ cambia mientras τ permanece fijo. Dependiendo de las condiciones en la frontera, las cuerdas pueden ser cerradas o abiertas (ver fig. 4). A medida que τ cambia, las cuerdas barren una superficie 2D llamada *hoja de mundo*, en la que muchas de sus propiedades tienen una descripción precisa (mediante una teoría *conforme*[10] de campo).

Para entender un poco mejor la descripción en la hoja de mundo de cuerdas habitando en nuestro universo, es útil comenzar con mecánica clásica. Una cuerda de longitud ℓ puede ser modelada por un continuo de resortes de masa m y constante k. Llamemos σ a la coordenada espacial que parametriza la curva descrita por la cuerda, y $X = X(\sigma, \tau)$ a la amplitud del estiramiento del resorte asociado al punto σ al tiempo τ. Dado que el Lagrangiano de un resorte con estas propiedades es $L = \frac{1}{2}m\,(\partial_\tau X)^2 - \frac{1}{2}k\,(\partial_\sigma X)^2$, para nuestro sistema obtenemos

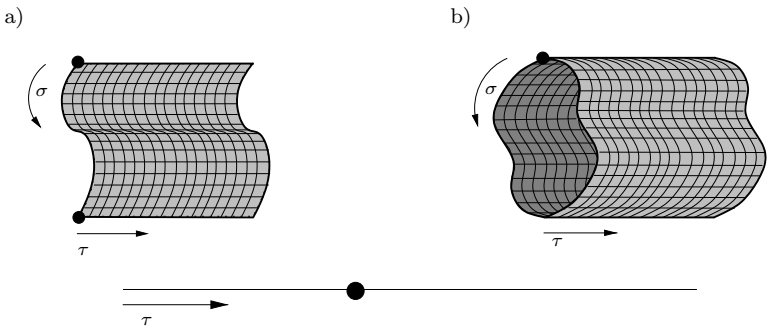


Figure 4: La línea de mundo de una partícula puntual es reemplazada por la hoja de mundo 2D barrida por un cuerda a) abierta o b) cerrada.

$$L = \int_0^\ell d\sigma \mathcal{L} = \frac{1}{2}\int_0^\ell d\sigma \left[m(\partial_\tau X)^2 - k(\partial_\sigma X)^2\right] = \frac{T}{2}\int_0^\ell d\sigma\, \partial^\alpha X \partial_\alpha X\,, \tag{21}$$

en donde $\alpha = 0, 1$ se refiere respectivamente a las coordenadas τ, σ. En la última expresión, hemos aplicado dos pasos. Primero, asumimos que m y k son iguales a la tensión T de la cuerda. Y, segundo, consideramos que la métrica del espacio $\tau - \sigma$ es Lorentziana, i.e. es la matriz 2D diagonal $h_{\alpha\beta} = \mathrm{diag}(1, -1)$.

Es útil comparar la densidad Lagrangiana de nuestra cuerda, $\mathcal{L} = \frac{T}{2}\partial^\alpha X \partial_\alpha X$, en (21) con $\mathcal{L}_{Higgs}$ en (19). Ambas expresiones se asemejan bastante en ausencia de energía potencial V. De hecho, $\mathcal{L}$ tiene la forma de la Lagrangiana para un campo escalar real libre (sin potencial), en un ambiente 2D carente de interacciones de norma (en este caso, la derivada covariante coincide con la parcial). Esto sugiere que la descripción de una cuerda puede entenderse como una QFT bidimensional en la que hay un campo bosónico libre $X(\sigma, \tau)$. Sin embargo, para poder establecer esta conexión, es preciso cuantizar el sistema 2D, considerando que las simetrías clásicas de $\mathcal{L}$ no deben de alterarse en la teoría cuántica (i.e. no debe haber *anomalías*). Este es el trabajo realizado hace cerca de 40 años [15–17]. Para que esta teoría no tuviera ninguna anomalía, se descubrió que se deben incluir 26 campos bosónicos X^M, tal que la Lagrangiana adopta la forma

$$\mathcal{L} = \frac{T}{2}\partial^\alpha X^M \partial_\alpha X_M\,, \qquad M = 0, \ldots, 25\,. \tag{22}$$

Regresando a nuestro estudio de la cuerda clásica, recordamos que X describía la amplitud de oscilación de la cuerda en una dirección de *nuestro* espacio-tiempo. Entonces, la predicción de la teoría de cuerdas que hemos descrito hasta ahora es que el espacio-tiempo tiene 26 direcciones independientes, o bien, 26 dimensiones.

[10] Este término hace referencia a teorías invariantes bajo un cierto grupo de simetrías; sin embargo, para la teoría de cuerdas considerada aquí, sólo se refiere a una teoría invariante bajo reescalamientos, i.e. donde no existen escalas fijas de ninguna índole.

Es importante en este punto remarcar la esencia de lo aprendido hasta aquí. Al escribir el Lagrangiano clásico de una cuerda vibrante, notamos la enorme similitud de éste con el que describe la dinámica de un campo escalar real en el espacio bidimensional llamado hoja de mundo. El campo escalar representa clásicamente la dirección en la que la cuerda vibra. Esto sugiere que, de ser posible cuantizar consistentemente la cuerda, la teoría resultante será una teoría de campos en la hoja de mundo 2D, dando origen a una teoría con espacio-tiempo unidimensional. El ejercicio de cuantizar esta cuerda (*bosónica*) preservando las simetrías clásicas de $\mathcal{L}$ en 2D requiere 26 campos escalares y, por lo tanto, describe un universo con 22 dimensiones extra.

Las ecuaciones de movimiento que surgen a partir de la Lagrangiana (22) usando el formalismo de Euler-Lagrange (5) son muy parecidas a las ecuaciones del electromagnetismo, ec. (3):

$$\partial_\alpha \partial^\alpha X^M \;=\; 0\,. \tag{23}$$

I.e. como era de esperarse, cada X^M se comporta como una onda. La solución de (23) es una serie de Fourier compuesta por *modos vibracionales* que se desplazan hacia un lado (digamos, a la derecha) y otros desplazándose hacia el lado opuesto (a la izquierda) a la velocidad de la luz[11]:

$$X^M = f_R(\sigma-\tau) + f_L(\sigma+\tau) = x^M + p^M\tau + \sum_{n\neq 0}\left(a_n^M e^{\pi i n(\sigma-\tau)/\ell} + \tilde{a}_n^M e^{-\pi i n(\sigma+\tau)/\ell}\right)\,, \tag{24}$$

en donde a_n^M y $\tilde{a}_n^M$ son coeficientes derechos e izquierdos respectivamente, y x^M y p^M denotan la posición y el momento del centro de masa de la cuerda. Para cuerdas cerradas, los modos derechos f_R y los modos izquierdos f_L son independientes.

La solución (24) es válida también cuánticamente. De hecho, esta ecuación puede interpretarse como la superposición de una infinidad de osciladores harmónicos desacoplados, lo cual se vuelve evidente al imponer las relaciones canónicas de conmutación entre X^M y el momento asociado $P^M = \partial\mathcal{L}/\partial(\partial_\tau X^M) = T\partial_\tau X^M$ (con $\hbar = 1$):

$$[X^M(\sigma),\, P^N(\sigma')] = -i\delta(\sigma-\sigma')\eta^{MN} \quad\Rightarrow\quad [a_n^M,\,(a_m^N)^\dagger] = [\tilde{a}_n^M,\,(\tilde{a}_m^N)^\dagger] \propto \eta^{MN}\delta_{m,n}\,, \tag{25}$$

en donde $n>0$, $(a_n)^\dagger \propto a_{-n}$ (y análogamente para $\tilde{a}_n$) y $\eta^{MN} = \mathrm{diag}(1,-1,\ldots,-1)$ es la métrica Lorentziana en 26 dimensiones. En el contexto cuántico, entonces, los coeficientes $a_n^\dagger, \tilde{a}_n^\dagger$ $(a_n, \tilde{a}_n)$ juegan el papel de operadores de creación (aniquilación), actuando sobre el vacío del espacio-tiempo 26D. La acción de los operadores de creación sobre el vacío genera el espacio de Hilbert de los estados que llenan el espacio-tiempo. A distancias $\gg \ell_{str} \sim 10^{-33}$ cm, la estructura de las cuerdas es imperceptible y, por tanto, los estados de Hilbert son percibidos como campos que describen partículas puntuales con propiedades familiares, tales como masa y espín.

Rápidamente, en los inicios de la teoría de cuerdas [17], se calculó el *espectro* (i.e. el conjunto de todos los estados de Hilbert) de la *cuerda bosónica* estudiada hasta aquí y se hicieron tres observaciones muy importantes: i) todos los estados descritos son bosones, ii) existe un bosón con masa cuadrada negativa, y iii) existe un bosón con $s=2$. En la época en la que se esperaba que la teoría de cuerdas funcionara como una descripción de las interacciones fuertes, estas observaciones fueron fatales. Como hemos visto, el SM contiene en su mayor parte fermiones y son los fermiones los que juegan un papel importante en la dinámica de la física nuclear. Además, el estado de masa cuadrada negativa (i.e. un taquión) representa una inestabilidad del vacío de la teoría, o sea que el vacío decae. Estas dos objeciones pueden evitarse en el contexto de la teoría de cuerdas invocando una simetría adicional que explicaremos más abajo. Sin embargo, la última objeción es infranqueable: ninguna partícula elemental relacionada con las interacciones fuertes tiene espín 2.

[11] Recordemos que, en unidades naturales –usadas en estas notas–, $c=1$.

2.1.2 La primera revolución

Sin embargo, lo que representó la estaca para la teoría de cuerdas como una posible descripción de la fuerza fuerte, fue lo que condujo, una década más tarde, a la denominada *primera revolución* de la teoría de cuerdas. El punto clave es que, si la gravedad logra ser cuantizada, debido a que ésta es una fuerza siempre atractiva y que está relacionada con la métrica del espacio-tiempo, la partícula mediadora de las interacciones gravitacionales llamada gravitón debe tener $s = 2$. En la primera revolución de la teoría de cuerdas, se demostró que ésta no sólo incluye una partícula con las propiedades del gravitón, sino que, además, la acción del gravitón es, en una primera aproximación a bajas energías, la acción de Einstein-Hilbert en 26 dimensiones. Por si fuera poco, debido a la naturaleza extendida de las cuerdas, desaparecen las divergencias ultravioletas, usualmente presentes en otros intentos de gravedad cuántica en los que las partículas son puntuales.[12] Por lo tanto, uno de los resultados más sorprendentes de la teoría de cuerdas es que provee un ambiente propicio para hablar de gravedad cuántica [17].

Supersimetría. A pesar de esta magnífica noticia, no muchos se convencieron de que una teoría que exhibe sólo bosones y, entre ellos, un taquión, tuviera alguna relevancia física. Sin embargo, poco tiempo después llegó la solución. La respuesta es la llamada *supersimetría* (SUSY). Esta es la máxima extensión de la simetría de Poincaré, que, de existir, prescribe que la física del universo es invariante bajo el intercambio de bosones por fermiones, y viceversa. Como consecuencia, para cada bosón de una teoría, debe existir su supercompañero fermiónico. La pregunta que se planteó entonces es ¿qué pasa si SUSY existe en la hoja de mundo? ¿Cómo afecta esto a la física del espacio-tiempo?

La tarea consistió en incluir, para cada campo bosónico X^M en $\mathcal{L}$ su correspondiente compañero supersimétrico Ψ^M. Esto cambia también las condiciones sobre la inexistencia de anomalías en la versión cuántica. En partícular, en lugar de necesitar 26 campos X^M, como en la cuerda bosónica, se requieren 10 parejas (X^M, Ψ^M) para evitar inconsistencias cuánticas. Es decir, las cuerdas supersimétricas predicen un universo 10-dimensional [18, 19].

Por completez, escribamos también en este caso la correspondiente densidad Lagrangiana. Usando lo aprendido sobre la Lagrangiana para el electromagnetismo (ver $\mathcal{L}_e$ en la ec. (7)), podemos conjeturar que la correspondiente Lagrangiana en la hoja de mundo está dada por

$$\mathcal{L} = \frac{T}{2}\left(\partial^\alpha X^M \partial_\alpha X_M + i\overline{\Psi}^M \Gamma^\alpha \partial_\alpha \Psi_M\right), \qquad M = 0, \ldots, 9, \tag{26}$$

con las matrices bidimensionales de Dirac expresadas como

$$\Gamma^0 = \begin{pmatrix} 0 & \mathrm{i} \\ -\mathrm{i} & 0 \end{pmatrix} \qquad \Gamma^1 = \begin{pmatrix} 0 & \mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}. \tag{27}$$

En este caso, tras calcular y resolver las ecuaciones de Euler-Lagrange para todos los campos, se llega a que también Ψ^M puede expresarse como una serie, cuyos modos de oscilación se traducen en operadores de creación y aniquilación en la teoría cuántica. Como en la cuerda bosónica, los operadores de creación inducen estados físicos en el espacio-tiempo 10D.

[12] Históricamente, el mayor de los problemas al cuantizar gravedad es que, al calcular la magnitud de las interacciones gravitacionales (que involucran dos o más lazos) a muy cortas distancias o, equivalentemente, a muy altas enegías (de aquí proviene el adjetivo 'ultravioleta'), aparecen integrales que divergen. Este problema está asociado a interacciones puntuales que no ocurren en teoría de cuerdas porque las interacciones suceden en las superficies que describen la colisión de las cuerdas.

Es posible calcular el espectro de cuerdas supersimétricas, lo que conduce de manera general a las siguientes propiedades: i) la supersimetría es "heredada" de la hoja de mundo al espacio-tiempo, i.e. hay bosones y fermiones en 10D también, ii) no existen taquiones, y iii) el gravitón con su supercompañero, el *gravitino*, forman parte del espectro. Este espectro de estados representa cierto progreso con respecto a QFT, ya que evita los problemas de la cuerda bosónica e incluye la gravedad de manera cuántica. Pero hay más: además de estos estados, existen otros que dependen de las características de las cuerdas bajo consideración, e.g. de si son cerradas o abiertas, si se transforman quiralmente o no.

Las cuerdas supersimétricas se dividen en tipo I y tipo II para cuerdas supersimétricas abiertas y cerradas, respectivamente. La ventaja de las cuerdas tipo I frente a las tipo II es que, como las cuerdas son dinámicas, las cuerdas abiertas del espectro pueden ocasionalmente cerrarse, conduciendo a estados presentes en el espectro de las cuerdas tipo II. Además, la consistencia de la teoría tipo I demanda la existencia del grupo de norma SO(32) con sus respectivos 496 bosones de norma (y los correspondientes supercompañeros fermiónicos llamados *gauginos*). En cambio, en las cuerdas tipo II sólo hay cuerdas cerradas y, con los ingredientes aquí descritos, no aparecen bosones de norma.

Existe un tipo adicional de supercuerdas. Como mencionamos antes, para las cuerdas cerradas los modos de vibración derechos son independientes de los modos de vibración izquierdos. Esta propiedad fue tomada como ventaja para construir un tipo híbrido de supercuerdas llamadas *cuerdas heteróticas* [20, 21].[13] Estas cuerdas son el resultado de combinar una parte de la cuerda tipo II (digamos, la derecha) con una parte de la cuerda bosónica (la izquierda). Es la diferencia en dimensionalidad (26D de la bosónica contra 10D de la tipo II) lo que da 16 grados de libertad bosónicos adicionales a la nueva cuerda supersimétrica 10-dimensional que conducen a los grupos de norma SO(32) y $E_8 \times E_8$. De esta manera, encontramos una posible explicación del origen de los grupos de norma en la física de partículas elementales: si la física observable proviene de las cuerdas tipo I o las heteróticas, los grupos de norma SO(32) o $E_8 \times E_8$ podrían ser los antecesores de las simetrías observadas.

Una propiedad relevante de las cuerdas aquí descritas es que las interacciones entre los campos emergentes de las teorías de cuerdas son proporcionales a una sola constante de acoplamiento g_s (y no tres, como en el SM), de tal manera que, si todas las partículas conocidas surgen de la teoría de cuerdas, entonces basta con medir un solo parámetro para caracterizar completamente las interacciones entre ellas. Esto conduce a un concepto muy popular en la física de partículas elementales: *unificación* de las fuerzas fundamentales. Si la teoría de cuerdas es la raíz de la física observable, entonces todas las fuerzas –con sus correspondientes constantes de acoplamiento g_i– deben surgir de una sola fuerza "madre", cuya magnitud está determinada por g_s.

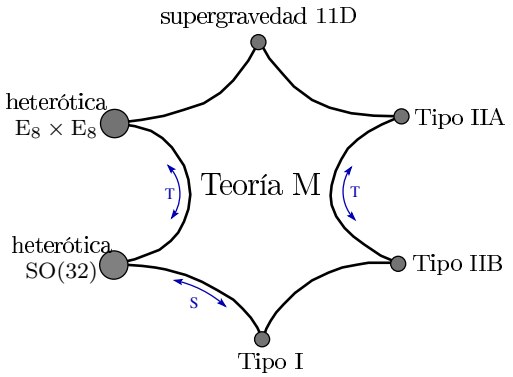


Figure 5: Las teorías de supercuerdas y sus conexiones.

En resumen, hemos visto que para lograr consistencia de la teoría de cuerdas es preciso exigir la existencia de una nueva simetría entre bosones y fermiones (SUSY) a energías tan altas como la escala energética de Planck ($M_{Pl} \sim 10^{18}$ GeV). Esta exigencia aunada a las condiciones que

[13] La palabra 'heterosis' en biología se refiere a la mezcla genética que conduce a especímenes con características más exitosas biológicamente.

permiten que las simetrías clásicas sean preservadas a nivel cuántico predicen que el espacio-tiempo en el que viven las partículas elementales (originadas por las vibraciones de las cuerdas) tiene 10 dimensiones. El espectro de las cuerdas supersimétricas incluyen, además del gravitón, fermiones y, en el caso de la cuerda tipo I y las heteróticas, bosones de norma junto con sus compañeros supersimétricos. Existen sólo 5 teorías de supercuerdas consistentes (ver fig. 5): la tipo I con simetría de norma SO(32), las tipo II (B y A, según tenga fermiones quirales o no), y las cuerdas heteróticas con simetrías de norma $E_8 \times E_8$ y SO(32).

Hagamos una breve reflexión acerca de los retos que estas teorías exhiben al intentar describir nuestro universo (abordaremos con más detalle estos tópicos en la sección 2.2). El mayor de los conflictos es que las cinco teorías de supercuerdas predicen un espacio-tiempo 10-dimensional mientras nuestra experiencia nos indica que existen sólo 4 dimensiones. La forma de evitar este problema es por medio de la *compactificación*, que consiste en considerar que (por alguna razón aún desconocida) las 6 dimensiones extra se distinguen de las 4 observadas en que las primeras son compactas[14] y, además, tan pequeñas que escapan a todo esfuerzo por detectarlas. Dado que la escala fundamental de las cuerdas es aproximadamente la escala de Planck, el hipervolumen del espacio extra mide $\sim (10^{-33}\ \text{cm})^6$, que escapa a todo experimento moderno. Otro problema es que nuestro universo no parece ser supersimétrico: e.g. nadie ha observado al compañero bosónico del electrón (llamado *selectrón*), ni al compañero fermiónico del fotón (llamado *fotino*). Este problema se soluciona típicamente al invocar un proceso parecido al producido por el mecanismo de Higgs, es decir, SUSY se rompe espontáneamente cuando algún campo supersimétrico adquiere un valor de expectación.

Por otra parte, las simetrías de norma que aparecen en las cuerdas heteróticas son inmensas. Tanto SO(32) como $E_8 \times E_8$ conducen a 496 bosones de norma, mientras que el SM tiene sólo 12. El método habitual para explicar esta discrepancia es nuevamente mediante la compactificación. En general, al compactificar, algunos de los bosones de norma adquieren masas del orden de M_{Pl}, por lo que, a bajas energías, sólo un subgrupo de los grupos de simetría originales es percibido. Finalmente, las cuerdas tipo II en esta primera revisión no contienen grupos de norma, contrario a lo que la física de partículas elementales requiere. Sin embargo, posteriormente se descubrió que la inclusión de hipersuperficies dentro del espacio 10-dimensional produce también otros grupos de norma en las teorías excluídas originalmente. A estos objetos se les llama *D-branas* y son parte importante del segundo gran movimiento "revolucionario" al interior de la teoría de cuerdas.

2.1.3 La segunda revolución

La llamada *segunda revolución* de la teoría de cuerdas ocurrió hace 3 lustros y vió nacer muchas nuevas formas de discutir la física moderna basadas en las herramientas de la teoría. Los ingredientes fundamentales de esta nueva ola de progreso en el campo fueron: i) Dualidades y ii) D-branas.

Dualidades. La mayor contribución de la segunda revolución de la teoría de cuerdas es que proporcionó una forma de incluir efectos no-perturbativos en la teoría, es decir, efectos similares a los de las interacciones fuertes a bajas energías, en los que expansiones en el acoplamiento (g_3 para la fuerza fuerte) no son válidas porque éste no es $\ll 1$. En general, la única clase de teorías que sabemos manipular son teorías perturbativas, en las que expansiones en serie son válidas. La pieza clave para resolver esta cuestión son las dualidades. De manera esquemática, una dualidad puede

[14] Por ejemplo, una dimensión compacta podría ser el círculo S^1 de radio R.

entenderse con el siguiente diagrama:

Teoría A $\overset{aprox.A}{\longrightarrow}$ Teoría emergente $\overset{aprox.B}{\longleftarrow}$ **Teoría B**

Se dice que una "Teoría A" es dual a la "Teoría B" si existe una aproximación en cada una de las teorías que conduce independientemente al mismo resultado, al que podemos llamar la "Teoría emergente".

En teoría de cuerdas, se encontró que hay dos dualidades esenciales, capaces de relacionar a todas las teorías: dualidad T y dualidad S. La teoría A es T-dual a la teoría B si la física resultante al considerar que una de las dimensiones de la Teoría A está *compactificada* en el círculo de radio R es equivalente a la física resultante al compactificar una dimensión de la teoría B en un círculo de radio $1/R$. (Una discusión interesante al respecto se puede encontrar en [22].) Esto quiere decir que, si las dimensiones extra son compactas, los efectos en la física de una teoría cuyo espacio compacto tiene volumen muy pequeño (e.g. $(10^{-33}\text{ cm})^6$) son equivalentes al resultado de hacer las dimensiones de la teoría T-dual muy grandes. Esta conexión es válida entre las teorías tipo II, la tipo IIA es T-dual a la teoría IIB; lo mismo ocurre entre las dos cuerdas heteróticas. Esto reduce el número de teorías de cuerdas independientes de 5 a 3.

Una implicación de este mapeo entre teorías T-duales es que los conceptos geómetricos usuales carecen de sentido y deben reemplazarse por un nuevo tipo de "geometría cuántica", que es descrita matemáticamente por una teoría conforme de campos bidimensional.

Supongamos ahora que la teoría A y teoría B son S-duales. Esto significa que, si $\mathcal{O}$ denota una cierta observable y g es la constante de acoplamiento en las interacciones entre los distintos campos, entonces

$$\mathcal{O}_A(g) \;=\; \mathcal{O}_B(1/g)\,. \tag{28}$$

Esta dualidad, cuya identificación [23] fue el primer paso de la segunda revolución de la teoría de cuerdas, generaliza la simetría electricidad-magnetismo de la teoría de Maxwell. En la teoría de Maxwell cuantizada (o sea, QED) la unidad básica de carga magnética q_m es inversamente proporcional a la unidad de carga eléctrica q_e que, como vimos antes, determina la magnitud de las interacciones electromagnéticas. Por lo tanto, la electricidad es S-dual del magnetismo. En la teoría de cuerdas, la dualidad S relaciona la cuerda tipo I con la teoría heterótica SO(32) y la teoría tipo IIB consigo misma. Una interpretación útil de la dualidad S es que una teoría acoplada fuertemente (con $g > 1$) que no puede ser estudiada perturbativamente, puede entenderse estudiante la teoría S-dual con acoplamiento débil (con $g < 1$). Esta interpretación es bastante útil para completar en el entendimiento de las teorías. Por ejemplo, la parte no-perturbativa de la teoría heterótica SO(32) puede entenderse completamente mediante el estudio *perturbativo* de la teoría I.

Es precisamente, la búsqueda de la explicación del comportamiento no-perturbativo de las teorías de cuerdas lo que condujo eventualmente al descubrimiento de la *teoría M*, una teoría 11-dimensional que, en el límite en el que una de sus dimensiones es compacta, conduce a las teoría de cuerdas heterótica $E_8 \times E_8$ y la tipo IIA. De esta manera, todas las teorías pueden relacionarse mediante una combinación de dualidad S, dualidad T y compactificaciones de la teoría M. Esta situación está representada en la fig. 5.

D-branas. En los 90's, J. Polchinski descubrió que la teoría de cuerdas requiere la inclusión de objetos de mayor dimensionalidad que las cuerdas (unidimensionales), llamadas Dirichlet p-branas o simplemente D-branas [11, 12]. El nombre se deriva de las condiciones a la frontera asignadas a los extremos de cuerdas abiertas. Las cuerdas abiertas usuales de la cuerda tipo I satisfacen una condición (de borde tipo Neumann) que asegura que el momento no fluye hacia o del extremo de la

cuerda. Sin embargo, la dualidad T implica la existencia de cuerdas abiertas duales con posiciones definidas (condiciones de borde tipo Dirichlet) en las dimensiones que son T-transformadas. En general, en las cuerdas tipo II, se pueden considerar cuerdas abiertas con posiciones fijas para los extremos en algunas dimensiones, lo que implica que las cuerdas son forzadas a terminar en una superficie preferida. A primer vista, esto aparenta romper la invariancia relativista de la teoría. Esta paradoja se resuelve al introducir hipersuperficies p-dimensionales en las que las cuerdas abiertas tienen sus extremos.

Se ha probado que objetos de mayor dimensionalidad que las cuerdas inducen efectos no-perturbativos en la teoría. Uno de los aspectos relevantes de las D-branas surge del hecho de que es posible estudiar sus excitaciones usando la misma QFT bidimensional perturbativa (de la hoja de mundo) de las cuerdas abiertas que hemos descrito antes, en lugar de una teoría de campos no-renormalizable sobre el (hiper)volumen de mundo de la D-brana misma. De esta manera, nuevamente encontramos una forma de estudiar efectos no-perturbativos mediante métodos perturbativos (expansiones en series). Es necesario remarcar aquí que las D-branas, así como las cuerdas mismas, son objetos dinámicos, es decir, cambian en el tiempo, lo que puede conducir a efectos no-perturbativos de relevancia cosmológica (como la aparición de agujeros negros).

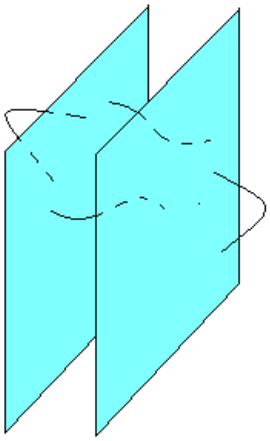


Figure 6: Una pila de 2 D-branas. Las cuerdas abiertas independientes constituyen los 4 bosones del grupo de norma $\mathrm{U}(2) = \mathrm{SU}(2) \times \mathrm{U}(1)$.

Quizá el aspecto más relevante de las D-branas es que conducen a simetrías de norma. Las simetrías de norma nacen al considerar todas las cuerdas abiertas independientes que pueden acoplar sus extremos a una D-brana o a un paquete de D-branas sobrepuestas. Las cuerdas abiertas consideradas aquí están orientadas, es decir, una cuerda orientada en una dirección no es equivalente a otra orientada en la dirección opuesta. Así, por ejemplo, en el caso de una *pila* de 2 D-branas sobrepuestas (cuya separación es nula), existen cuatro cuerdas independientes (ver fig. 6). Dicho de otra manera, para una pila de 2 D-branas, cada extremo de una cuerda abierta puede acoplarse a cualquiera de las 2 D-branas indistintamente, lo que representa 2^2 grados de libertad. En general, para pilas de N branas sobrepuestas, existen N^2 grados de libertad asociados a las cuerdas abiertas acopladas a la pila. La observación crucial es que los campos asociados a las cuerdas abiertas tienen las mismas propiedas que el campo del fotón y, por lo tanto, corresponden a bosones de norma. Dado que hay N^2 bosones de norma en sistemas de esta clase, el grupo de norma emergente es $\mathrm{U}(N)$, el cual es equivalente a $\mathrm{SU}(N) \times \mathrm{U}(1)$. Esto podría explicar por qué los grupos de norma que aparecen en la física de partículas son precisamente de la forma $\mathrm{SU}(N)$ (pudiendo ser $\mathrm{SO}(N)$ o cualquier otro).

Las D-branas añadieron una rica estructura matemática a la teoría y, como veremos, abrieron la puerta a muchas posibilidades para construir modelos con propiedades cercanas a las de la física observable.

2.2 Fenomenología de cuerdas

2.2.1 ¿Qué es fenomenología de cuerdas?

En la actualidad, existen dos visiones de la teoría de cuerdas: las podemos llamar la visión "clásica" y la visión "moderna". Éstas persiguen finalidades un poco diferentes que, en ocasiones, son consideradas ortogonales y están basadas en métodos matemáticos que difieren frecuentemente. Sin

embargo, ambas visiones estudian diversos aspectos de nuestro universo 4-dimensional mediante la aplicación de la magnífica estructura que la teoría de cuerdas ofrece. En la visión clásica, la teoría de cuerdas representa el origen de toda la física conocida. Es decir, una vez que todos los detalles de la teoría de cuerdas sean entendidos, se podrá identificar un modelo preferido que conduce a las interacciones del SM y al modelo cosmológico, dando respuesta a todas las interrogantes que estos admiten. En la visión clásica, pues, la teoría de cuerdas es la teoría de todo, la teoría final. Con este espíritu, se construyen modelos basados con las herramientas disponibles en la teoría (branas, compactificaciones, dualidades,...) que reproduzcan las características de nuestro universo y/o que den solución a los problemas sin resolver de la física moderna. La apuesta es que, eventualmente, se identificará el modelo ideal y, al mismo tiempo, el proceso dinámico que reduce la teoría de cuerdas a ese modelo.

El propósito en la visión clásica de la fenomenología de cuerdas es, pues, establecer lazos entre la teoría de cuerdas y las teorías bien fundamentadas experimentalmente a bajas energías. Así, se considera que la física de partículas elementales y macroscópica *emerge* de la teoría fundamental (como la mecánica clásica debe surgir de la mecánica cuántica) y que esta teoría fundamental es la teoría de cuerdas.

En la visión moderna, se abandona la ambiciosa tarea de la búsqueda de una teoría fundamental que dé la explicación de las preguntas más profundas de la física contemporánea. En su lugar, se usa a la teoría de cuerdas como una fuente de herramientas matemáticas que, empleadas dentro o fuera del contexto de la teoría de cuerdas, puedan resolver algún problema de la física actual. El máximo exponente de esta visión es posiblemente la dualidad norma/gravedad o correspondencia AdS/CFT o dualidad de Maldacena [24]. Descubierta como una aplicación del principio holográfico usual en discusiones sobre agujeros negros, esta dualidad ha probado ser útil para describir sistemas con interacciones fuertes, tales como superconductores, agujeros negros, el plasma de quarks y gluones, movimiento Browniano, entre otros. Se recomienda [25] como una revisión más detallada y útil para entender esta dualidad.

En estas notas, nosotros adoptamos la visión clásica. Partiendo de la teoría de cuerdas, describiremos algunos de los métodos conocidos que pueden conducir al SM, dejando de lado (para un paso posterior) los aspectos cosmológicos que, a pesar de ser omitidos en nuestra discusión, también están incluidos en los modelos aquí presentados. En el camino hacia el SM, sin embargo, hay una barrera que deberemos cruzar. Por construcción, el espacio-tiempo de una teoría de cuerdas in-

Teoría de cuerdas		**Fenomenología**		**Nuestro universo**
10 dimensiones $\mathbb{M}^{10}$		$\mathbb{M}^{10} \to \mathbb{M}^4 \times \cancel{X_6}$		4 dimensiones $\mathbb{M}^4$
SUSY		SUSY $\to$ ~~SUSY~~		~~SUSY~~
grupos de norma	$\Longrightarrow$	✓	$\Longrightarrow$	$\mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$
1 acoplamiento g_s		$g_s \to g_Y, g_2, g_3$		3 acoplamientos g_Y, g_2, g_3
$E \geq M_{Pl} \sim 10^{18}$ GeV		teoría efectiva		$E \leq M_{EW} \sim 10^2$ GeV
cuerdas				campos
gravitón+campos de norma		↓		campos de norma+quarks+leptones
gravedad cuántica		paso intermedio		gravedad clásica
estructura $\to$ simetrías		**MSSM** en 4D?		simetrías *ad hoc*

Table 2: Visión clásica de la fenomenología de cuerdas. La fenomenología de cuerdas establece el vínculo que relaciona a la física conocida a bajas energías con la teoría de cuerdas. En este contexto, la física de bajas energías es un límite *efectivo* de una teoría más *fundamental*, así como la mecánica clásica es considerada el límite macroscópico (de más bajas energías) de la mecánica cuántica.

cluye SUSY, i.e. incluye una simetría que requiere que todo fermión cuente con un supercompañero bosónico y, viceversa, que exista un fermión para todo bosón de la teoría. Si las cuerdas describen el SM, seguramente existen los supercompañeros de todas las partículas de la tabla 1. La versión completa del SM es llamada *MSSM* y se trata de una teoría de partículas supersimétrica. Es sensato, entonces, comenzar nuestra búsqueda de la física observable como una búsqueda del MSSM en el contexto de la teoría de cuerdas. En la sección 2.2.2, estudiaremos las características del MSSM.

En la tabla 2 esbozamos las metodología de fenomenología de cuerdas que seguimos. La primera columna describe las propiedades de la teoría de cuerdas discutidas en la sec. 2.1.2: espacio-tiempo 10-dimensional con métrica Lorentziana denotado como $\mathbb{M}^{10}$; SUSY tanto en la hoja de mundo como en el espacio-tiempo; (enormes) grupos de norma; 1 solo acoplamiento que determina la amplitud de las interacciones entre cuerdas; validez de la teoría a energías tan altas como la escala energética de Planck; los objetos elementales no son partículas sino cuerdas; e incluye simetrías de norma para todas las fuerzas fundamentales sin despreciar la gravedad, además de otras simetrías que provienen directamente de la estructura del espacio-tiempo. En la tercera columna, a manera de comparación, describimos las propiedades esenciales de la física conocida, como fue descrita en la sección 1.2: espacio-tiempo 4-dimensional tipo Minkowski $\mathbb{M}^4$; no hay SUSY o, en el mejor de los casos, la supersimetría está rota, ~~SUSY~~; grupos de norma (pequeños) para tres interacciones fundamentales; 3 acoplamientos g_i que determinan la magnitud de las interacciones de norma; validez de la teoría hasta las energías actualmente corroboradas en los aceleradores de partículas; los objetos elementales son campos que describen partículas; incluye a todas las partículas observadas (quarks y leptones) con sus respectivos mediadores de 3 de las 4 interacciones; sin embargo, la gravedad no es incluida en el tratamiento cuántico y todas las simetrías consideradas son puestas *a mano*.

En la segunda columna de la tabla 2 planteamos la sugerencia de la teoría de cuerdas para tratar las discrepancias. Primero, 6 de las 10 dimensiones de $\mathbb{M}^{10}$ son *compactificadas* en una variedad compacta X_6, tal que

$$\mathbb{M}^{10} \;=\; \mathbb{M}^4 \times X_6\,, \tag{29}$$

en donde $\mathbb{M}^4$ describe la geometría del espacio-tiempo que nos es familiar. En general, la variedad X_6 debe tener un volumen < 0.1 mm, para poder escapar a las mediciones y una estructura que, por motivos explicados en sección 2.2.2, preserva SUSY en 4 dimensiones. El rompimiento de la SUSY remanente requiere de la existencia de un *sector oculto*[15] en el que SUSY es rota de forma similar a como $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ es rota mediante el mecanismo de Higgs, i.e. cuando, dinámicamente, un campo supersimétrico adquiere un valor de expectación en el vacío. Es necesario concebir un mecanismo de transmisión o mediación de este rompimiento "oculto" de SUSY hacia el sector observable. En general, al menos la gravedad establecerá dicha mediación. Dado que la teoría de cuerdas da origen a grupos de norma, la tarea es concebir una compactificación con elementos tales que permitan la aparición del grupo de norma del SM, permitiendo así, a partir del valor de la constante de acoplamiento de cuerdas g_s derivar los valores de las constantes de acoplamiento de las tres fuerzas fundamentales del SM.

2.2.2 El modelo mínimo supersimétrico MSSM

Como mencionamos antes, entre la teoría de (super)cuerdas y el SM hay un paso intermedio: la mínima extensión supersimétrica del SM (MSSM, por sus siglas en inglés). La única diferencia entre

[15] Un "sector oculto" de la teoría es aquél cuyas partículas y fuerzas sólo interactúan gravitacionalmente con las partículas y fuerzas del MSSM que forman el *sector observable*.

el SM y el MSSM es que se incluyen los compañeros supersimétricos para todas las partículas de la tabla 1. La nomenclatura supersimétrica es la siguiente: el compañero bosónico de un fermión ψ del SM tiene el mismo nombre con el prefijo $s-$ y es denotado $\widetilde{\psi}$. Por ejemplo, los compañeros del electrón y del quark top son, respectivamente, el *selectrón* $\widetilde{e}$ y el *stop* $\widetilde{t}$. El compañero fermiónico de un bosón φ del SM añade el sufijo $-ino$ al nombre original y se denota como $\widetilde{\varphi}$; por ejemplo, $\widetilde{H}$ es un *Higgsino* y $\widetilde{W}$ es un *wino*. La sola excepción a estas reglas es el *gravitino* $\widetilde{\chi}$ que es el compañero del gravitón $g_{\mu\nu}$. Sin embargo, no será considerado en adelante porque no forma parte del MSSM. Todos las nuevas partículas supersimetricas tienen los mismos números cuánticos que sus compañeros presentes en el SM, i.e. los especificados en la tabla 1, salvo por el espín. Los supercompañeros de los quarks y leptones, i.e. los squarks y sleptons, tienen $s = 0$, mientras que los gauginos (compañeros de los bosones de norma) tienen spin $s = 1/2$. Los Higssinos también tienen $s = 1/2$.

El MSSM, como lo hemos definido arriba tiene un problema. Debido a la estructura de supersimetría, un sólo *supercampo* de Higgs no puede participar en las interacciones de Yukawa (ver ec. (20)) para ambos quarks u y d (y sus copias más pesadas). Además, si se incluye un sólo Higgsino, éste destruye la cancelación de anomalías de norma. Para resolver ambos conflictos es necesario añadir un segundo bosón de Higgs H' con su respectivo Higgsino. El segundo Higgs difiere del Higgs original en su hipercarga, $q_Y(H') = +\frac{1}{2}$.

Un ingrediente adicional en el MSSM es la llamada paridad R. La paridad R es una simetría $\mathbb{Z}_2$ que asigna carga $+1$ a todas las partículas del SM y carga -1 a sus supercompañeros. De esta manera, dado que la paridad R total en todo tipo de interacciones debe ser $+1$, los supercompañeros de las partículas del SM sólo interactúan en pares. Por ejemplo, mientras que la interacción fotino-selectrón-positrón $\widetilde{A}_\mu \widetilde{e}\bar{e}$ tiene paridad total $(-1)(-1)(+1) = +1$ y está permitida en el MSSM, la interacción fotino-electrón-positrón $\widetilde{A}_\mu e\bar{e}$ con paridad $(-1)(+1)(+1) = -1$ está prohibida. Una consecuencia inmediata es que la partícula supersimétrica más ligera no puede decaer en partículas del SM y, por lo tanto, es estable. Esto provee una partícula escalar (porque tiene $s = 0$) que podría explicar el origen de la matería oscura. La razón para añadir esta paridad es que, en su ausencia, la vida del protón sería de una fracción de segundo mientras que experimentos sitúan la vida media del protón como $> 10^{35}$ años.

El MSSM es la propuesta más elegante y simple para poder evitar el problema de jerarquía del SM explicado en la sección 1.2.2. Debido a la presencia de nuevas partículas (los supercompañeros), hay más contribuciones a la masa (o energía) del Higgs del SM. De manera contraintuitiva, las nuevas contribuciones cancelan las contribuciones de las partículas del SM, incluso cuando SUSY es ligeramente rota. Así, tenemos que la supersimetría, una simetría que aparece de manera natural en la teoría de cuerdas, parece ser un requisito fundamental para garantizar la consistencia de la física de partículas. Para algunos, la existencia de SUSY es una predicción de la teoría de cuerdas que probablemente será pronto confirmada en el acelerador LHC.

Así, pues, en un primer paso hacia la descripción de la naturaleza, la teoría de cuerdas debe proporcionar una estructura que contenga el MSSM.

2.2.3 Compactificaciones

Como hemos dicho antes, la forma canónica de explicar la indetectabilidad de las dimensiones extra predichas por la teoría de cuerdas es mediante el proceso de compactificación, matemáticamente establecido en ec. (29).

El espacio compacto 6-dimensional X_6 debe tener una estructura tal que no incluya inconsistencias en la teoría y que preserve SUSY en 4D, con la finalidad de llegar al MSSM, resolviendo el problema de jerarquía del SM. Con esta finalidad, una característica que debemos demandar de X_6

es *grupo de holonomía* SU(3) o un subgrupo de SU(3) más grande que SU(2).[16] Las variedades 6D que tienen esta propiedad y que además son suaves (contínuas) se llaman variedades de Calabi-Yau (CY) en honor a sus descubridores. La figura 7 presenta una proyección 3D de la variedad CY quíntica.

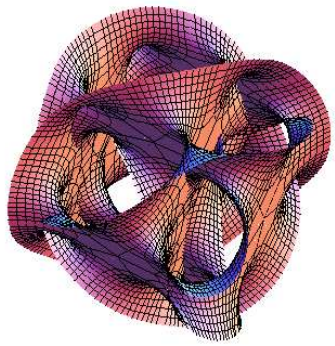


Figure 7: Proyección tridimensional de la variedad CY llamada quíntica.

La propuesta de solicitar que el espacio 6D compacto fuera una CY data de hace más de 20 años [26]. En la propuesta original, se compactificó el espacio extra de la cuerda heterótica con grupo de norma $E_8 \times E_8$ y se notó que, sin otros ingredientes, el grupo de norma en 4D es reducido a $E_6 \times E_8$. Si se considera que el segundo factor E_8 corresponde a un sector oculto, entonces el sector observable, con grupo E_6, contiene 4 **27**-pletes (análogos a los dobletes de $SU(2)_L$, pero con 27 partículas) que corresponden a 4 generaciones de quarks y leptones (más materia adicional) en el lenguaje de las teorías de gran unificación (ver e.g. [7] para mayores detalles sobre estas teorías). Este modelo tiene una serie de problemas: el grupo de norma del MSSM no es E_6, el MSSM tiene 3 generaciones y no 4, y no hay forma de evitar las partículas adicionales (e.g. el número de bosones de norma en E_6 es 78, mientras que el MSSM sólo tiene 12). Sin embargo, la enorme cualidad de este modelo es que sentó las bases de cómo se podría lograr una compactificación a 4D y romper la simetría de norma original al mismo tiempo.

Actualmente se sabe que no sólo las cuerdas heteróticas pueden compactificarse en variedades CY, sino que todas las variantes de la teoría de cuerdas admiten dicha compactificación. Además, también se sabe que es posible obtener exactamente los estados del MSSM [27] cargados bajo el grupo de norma $SU(3)_c \times SU(2)_L \times U(1)_Y \times U(1)_{B-L}$. La simetría de norma $U(1)_{B-L}$ (con B número hadrónico –de hecho, bariónico– y L número leptónico) no es una simetría de norma del MSSM. Aunque es un modelo atractivo, la incapacidad de liberarnos de la simetría de norma adicional motiva buscar nuevos modelos. Una desventaja de las compactificaciones tipo CY es que representan matemáticamente todo un reto. De hecho, incluso los matemáticos han sido incapaces de clasificar todas las variedades de este tipo. Por otra parte, hasta la fecha de estas notas, no existe un método práctico y confiable de calcular la magnitud de las interacciones entre los campos originados por la teoría de cuerdas, lo que le resta capacidad predictiva a los modelos. La pregunta es, entonces, si existe una forma más simple de llegar a mejores resultados que los descritos aquí.

La respuesta no se hizo esperar. Poco tiempo después de la propuesta de las compactificaciones en variedades CY, llegó una nueva sugerencia. Las compactificaciones de la teoría 10D en un orbifold 6D es, por excelencia, la forma más simple y prometedora de compactificar consistentemente cualquiera de los tipos de cuerdas [28, 29]. Un orbifold es el resultado de dividir una variedad compacta por una de sus simetrías discretas. Por ejemplo, supongamos que una teoría posee una sola dimensión adicional X, como propusieron Kaluza y Klein hace casi un siglo. Como se muestra en la figura 8, la dimensión adicional se puede "enrollar" para formar una variedad S^1 aplicando la identificación $X \simeq X + 2\pi R$, con R el radio que define el "tamaño" del espacio compacto. El orbifold 1D se logra al dividir una de las simetrías discretas de esta variedad. La única simetría disponible es una reflexión $X \simeq -X$, cuyo grupo es $\mathbb{Z}_2$ (porque al aplicarla dos veces equivale a la

[16] El grupo de holonomía se determina estudiando todos los vectores que resultan del transporte paralelo a lo largo de curvas cerradas sobre la superficie de una variedad compacta. Por ejemplo, el grupo de holonomía del círculo S^1 es trivial (porque, el transporte paralelo de un vector en curvas cerradas sobre S^1 conduce siempre a sí mismo), mientras que el grupo de holonomía de S^2 es el grupo de rotaciones en el plano 2D tangente a cada punto de la esfera, i.e. SO(2)

acción de la identidad).

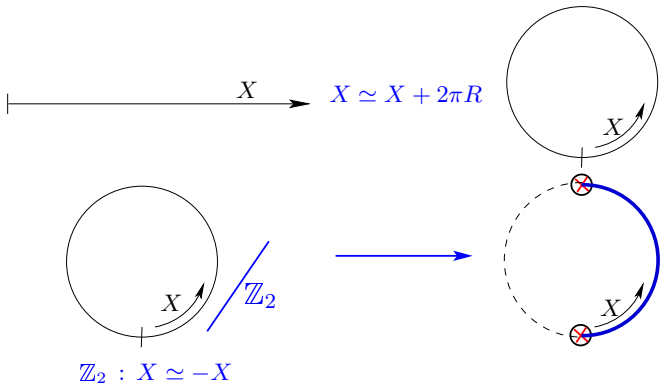


Figure 8: Orbifold 1-dimensional $\mathbb{Z}_2$. Una dimensión es compactificada en S^1 y entonces una simetría discreta $\mathbb{Z}_2$ es dividida.

El resultado de la compactificación de la dimensión adicional en el orbifold $\mathbb{Z}_2$ es que el *dominio fundamental* de la variedad, i.e. la descripción de la dimensión entera, se reduce a la mitad del dominio en S^1. Aunado a este efecto, surgen dos puntos que no son alterados por la reflexión $\mathbb{Z}_2$. Estos puntos fijos son singularidades de curvatura del orbifold, i.e. el escalar de curvatura diverge en esos puntos. Estas divergencias podrían parecer peligrosas, pero debemos recordar que habitan en el espacio compacto y no en el espacio-tiempo 4D $\mathbb{M}^4$, el cual es ortogonal a la dimensión compacta. En este caso, si $R < 0.1$ mm, la dimensión extra puede escapar al más riguroso intento de medición.[17]

Nuestro objetivo es aplicar este tipo de compactificaciones a las teorías de cuerdas. El procedimiento habitual consiste en compactificar las 6 dimensiones adicionales en un toro 6D factorizable, i.e. $T^6 = T^2 \times T^2 \times T^2$. El orbifold 6D se obtiene al dividir una simetría discreta de T^6. Por ejemplo, consideremos el orbifold $\mathbb{Z}_2 \times \mathbb{Z}_2$ de la figura 9. En este caso, una simetría $\mathbb{Z}_2$ es dividida a cada T^2, lo que quiere decir que, como en el caso 1D, sólo se efectúa la identificación $X^i \simeq -X^i$ con $i = 1, \ldots, 6$. En este caso, el dominio fundamental del espacio compacto se reduce a la mitad y produce 4 puntos invariantes (bajo $\mathbb{Z}_2 \times \mathbb{Z}_2$) en cada uno de los T^2. Los puntos fijos son singularidades de curvatura en el espacio compacto que pueden interpretarse como las esquinas de 3 tetrahedros ortogonales. El grupo de holonomía del orbifold 6D es, en general, un subgrupo discreto de SU(3) que no puede incorporarse en SU(2). Esto, como en el caso de las variedades CY, conduce a SUSY en 4D.

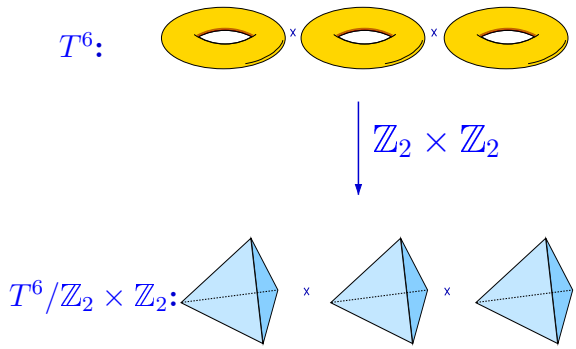


Figure 9: Orbifold 6-dimensional $\mathbb{Z}_2 \times \mathbb{Z}_2$. 6 dimensiones son compactificadas en $T^6 = T^2 \times T^2 \times T^2$ y entonces una simetría discreta $\mathbb{Z}_2 \times \mathbb{Z}_2$ es dividida.

Este procedimiento es particularmente útil en las cuerdas heteróticas, que es donde se aplicó por vez primera [28,29]. Dependiendo de los detalles geométricos de T^6 y la simetría discreta elegida para dividirlo, los grupos de norma originales, SO(32) o $E_8 \times E_8$, son rotos a subgrupos que sobreviven en 4D.

Recientemente, en el llamado *Minilandscape heterótico* [30, 31], que divide una simetría $\mathbb{Z}_6$ al toro T^6, se exploraron 10 millones de modelos en búsqueda de construcciones que reproduzcan el

[17] Una forma de verificar si existen dimensiones extra compactificadas es midiendo la validez de la teoría de Newton a pequeñas escalas. De acuerdo a la teoría de la gravedad, la potencia de la distancia en la fuerza gravitacional cambia ante la presencia de dimensiones extra, cuando la detección se realiza por debajo del tamaño de dicha dimensión.

MSSM. Se encontraron alrededor de 300 modelos con las siguientes propiedades:

- el grupo de norma del MSSM,
- paridad R para evitar el decaimiento rápido del protón [32],
- el espectro exacto (sin partículas adicionales) del MSSM,
- unificación de todas las constates de acoplamiento a altas energías,
- escala de rompimiento de SUSY consistente con la solución del problema de jerarquía [33],
- mecanismo de sube-y-baja para generar masas de los neutrinos [34],

y otras características, tales como la posibilidad de resolver el problema de CP fuerte, posibles extensiones del MSSM con partículas neutras adicionales (el NMSSM), simetrías que podrían asegurar la estabilidad del protón en toda la historia del universo, entre otras.

En estos modelos, las generaciones de quarks y leptones del MSSM provienen de cuerdas que se encuentran localizadas en las singularidades de curvatura del orbifold, proveyendo así una posible explicación de por qué existen 3 generaciones y no otro número arbitrario. Además, dado que los puntos fijos forman polihedros, las generaciones se someten a las simetrías discretas (y no abelianas) de estas topologías, lo que conduce a *simetrías de sabor* que podrían explicar la magnitud de las masas y la estructura de mezclas de los quarks y leptones.

Un detalle a destacar en estas construcciones es la estadística obtenida. Si bien, 300 modelos prometedores entre 10 millones no suena muy alentador, la situación en muchos otros arreglos de cuerdas es bastante menos halagadora. Por ejemplo, en los modelos con D-branas que estamos a punto de evaluar, no existe hasta hoy un solo modelo que contenga todas las características que acabamos de exponer. Sin embargo, es posible que los modelos con branas más prometedores estén aguardando ser descubiertos pronto. Además, nada nos dice que una buena estadística conduzca al modelo que describe la naturaleza.

2.2.4 Mundos Brana

Otra forma de llegar a modelos de cuerdas que originen el MSSM es conocida como *mundos brana*. En este tipo de modelos, comenzando con las teorías de cuerdas tipo II, el espacio compacto se puede considerar tan grande como una fracción de milímetro porque la física que afecta a nuestro universo se encuentra en la intersección de distintas D-branas y por lo tanto, nuestras observaciones no son sensibles al tamaño de todo el espacio-tiempo 10D. O sea, los mundos brana asumen que el análisis se puede realizar de manera local en donde las D-branas estén localizadas $\rightarrow$ la estructura local es más importante que la global.

En estos escenarios, pilas de D-branas de distinta dimensionalidad, como las descritas en la sección 2.1.3, se enredan alrededor de ciclos no contractibles en regiones especiales del espacio compacto. No es difícil imaginar en estas construcciones arreglos en los que exista un paquete de 3 D-branas, un paquete de 2 D-branas y un paquete de una D-brana que darían origen al grupo de norma $\mathrm{U}(3) \times \mathrm{U}(2) \times \mathrm{U}(1) \supset \mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$.

Lo difícil de estos modelos es incluir las partículas del MSSM. Estas son creadas en las intersecciones de las D-branas. Son las cuerdas abiertas que comunican a branas de distintas pilas de D-branas las responsables de la materia fermiónica del SM. Por ejemplo, de los números cuánticos de la tabla 1, sabemos que un quark u_L (y su supercompañero) se transforma como un **3** bajo $\mathrm{SU}(3)_c$ y como un **2** bajo $\mathrm{SU}(2)_L$. Para conseguir esta situación, se requiere que exista una cuerda abierta que comunique la pila de 3 D-branas con la pila de 2 D-branas.

Hace casi una década, se descubrió el primer modelo consistente con los ingredientes citados [36]. Como se muestra en la figura 10, la mínima configuración consistente requiere de 4 paquetes de branas, en lugar de 3. Los paquetes de branas se intersectan de forma tal que producen tanto los quarks como los leptones izquierdos y derechos de una generación de quarks y leptones en el MSSM.

Una vez obtenida una generación de quarks y leptones, la siguiente tarea no trivial es obtener la multiplicidad de 3. En la figura 11, usamos un ejemplo en el que el espacio que envuelven las D-branas es un toro.

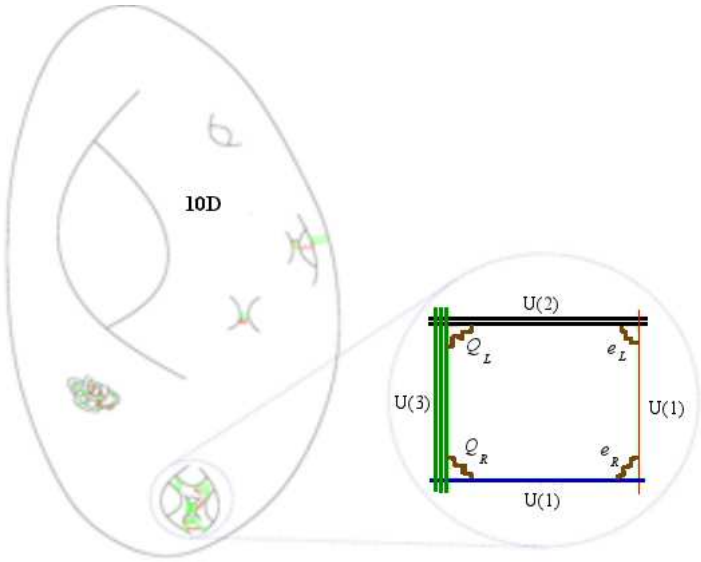


Figure 10: El SM como producto de D-branas en intersección. Los quarks y leptones (derechos e izquierdos) surgen en las intersecciones de un sistema con D-branas que conduce al grupo de norma $U(3)\times U(2)\times U(1)^2$. (Imagen cortesía de [35]).

En este caso, dependiendo del ángulo en el que las D-branas se intersecten, esta intersección puede ocurrir más de una vez. En la figura, hemos representado esquemáticamente la intersección de las branas que conducen a U(3) (en verde/horizontal) y las branas que conducen a U(2) (en negro). Las estrellas marcan el punto de intersección y también la localización de los 3 quarks izquierdos Q_L.

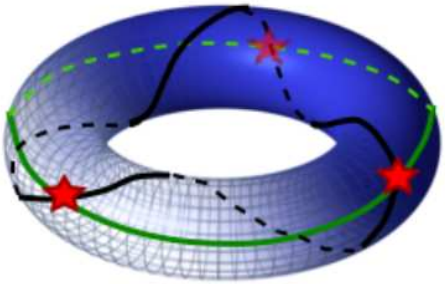


Figure 11: 3 generaciones como intersecciones en un toro. (Imagen cortesía de [35]).

Una tarea poco trivial en este tipo de modelos es la cancelación de anomalías. Dado que estas construcciones asumen un estudio local (en una pequeña región del espacio-tiempo) de las propiedades físicas del universo, se dejan de lado otros aspectos del mismo modelo que estarían sucediendo en otras regiones del espacio 10D. Esto provoca que, si bien las anomalías podrían ser canceladas localmente en la región en donde el SM aparece, también podría ocurrir que lejos de esta región aparecen nuevas contribuciones anómalas. Hoy se sabe que sí es posible logara la cancelación de anomalías en algunas construcciones particulares y que, al mismo tiempo, es posible encontrar modelos con la simetría de norma del MSSM y los campos requeridos por éste [37]. Lamentablemente, aún es controversial si existen además partículas exóticas.

En resumen, hemos visto en esta sección que existen métodos capaces de conducir a modelos que reproducen muchas cualidades del MSSM. Sin embargo, antes de comenzar a trazar posibles predicciones quedan muchas preguntas por resolver. En particular, la magnitud de las interacciones calculada con los métodos existentes en general conduce a funciones de ciertos campos escalares llamados *módulos*, cuyos valores en el vacío de la teoría no está fijos. Para poder reflejar los datos experimentales, es necesario comprender la naturaleza de los módulos y los procesos dinámicos que podrían conducir a un valor preferido. Además, es posible que, incluso tras lograr entender y

estabilizar los valores de los módulos, las predicciones no coincidan con los datos experimentales. De ser este el caso, el paso a seguir es continuar la búsqueda de modelos más prometedores, guiada por la experiencia adquirida en el paso anterior. Por supuesto, para cada modelo prometedor es imprescindible explorar todas sus propiedades y exigir que arroje datos que puedan ser corroborados experimentalmente. Estas son las tareas que la fenomenología de teoría de cuerdas tiene por delante y que deparan un sinnúmero de retos para los físicos de hoy.

Agradecimientos

Este trabajo ha sido parcialmente apoyado por el proyecto CONACyT 82291 y el proyecto DGAPA IA101811.

References

[1] Peskin M E and Schroeder D V *An Introduction to quantum field theory* reading, USA: Addison-Wesley (1995) 842 p.

[2] Einstein A 1909 *Physikalische Zeitschrift* **10** 817–825

[3] Debye P 1910 *Annalen der Physik* **338** 1427–1434

[4] Born, M and Heisenberg, W and Jordan, P 1926 *Zeitschrift für Physik A Hadrons and Nuclei* **35**(8) 557–615

[5] Dirac P A M 1927 *Royal Society of London Proceedings Series A* **114** 243–265

[6] Feynman R *QED: The strange theory of light and matter.* princeton University Press (1985). ISBN 0-691-08388-6

[7] Slansky R 1981 *Phys. Rept.* **79** 1–128

[8] Cahn R N 1984 *Semisimple Lie algebras and their representations* (Benjamin/cummings) 158 P. URL `http://phyweb.lbl.gov/~rncahn/www/liealgebras/book.html`

[9] Green M B, Schwarz J H and Witten E *Superstring Theory. Vol. 1: Introduction* cambridge, Uk: Univ. Pr. (1987) 469 P.

[10] Green M B, Schwarz J H and Witten E *Superstring Theory. Vol. 2: Loop amplitudes, anomalies and phenomenology* cambridge, Uk: Univ. Pr. (1987) 596 P.

[11] Polchinski J *String theory. Vol. 1: An introduction to the bosonic string* cambridge, UK: Univ. Pr. (1998) 402 p

[12] Polchinski J 1998 *String theory. Vol. 2: Superstring theory and beyond* (Cambridge, UK: Univ. Pr.) 531 P.

[13] Bailin D and Love A *Supersymmetric gauge field theory and string theory* bristol, UK: IOP (1994) 322 p. (Graduate student series in physics)

[14] Green B *The elegant universe.* vintage Books (2000) 448 p.

[15] Yoneya T 1973 *Nuovo Cim. Lett.* **8** 951–955

[16] Yoneya T 1974 *Prog. Theor. Phys.* **51** 1907–1920

[17] Scherk J and Schwarz J H 1974 *Nucl. Phys.* **B81** 118–144

[18] Gliozzi F, Scherk J and Olive D I 1977 *Nucl. Phys.* **B122** 253–290

[19] Brink L, Schwarz J H and Scherk J 1977 *Nucl. Phys.* **B121** 77

[20] Gross D J, Harvey J A, Martinec E J and Rohm R 1985 *Phys. Rev. Lett.* **54** 502–505

[21] Gross D J, Harvey J A, Martinec E J and Rohm R 1985 *Nucl. Phys.* **B256** 253

[22] Witten E 1996 *Phys. Today* **49N4** 24–30

[23] Font A, Ibáñez L E, Lüst D and Quevedo F 1990 *Phys. Lett.* **B249** 35–43

[24] Maldacena J M 1998 *Adv. Theor. Math. Phys.* **2** 231–252 (*Preprint* `hep-th/9711200`)

[25] Aharony O, Gubser S S, Maldacena J M, Ooguri H and Oz Y 2000 *Phys. Rept.* **323** 183–386 (*Preprint* `hep-th/9905111`)

[26] Candelas P, Horowitz G T, Strominger A and Witten E 1985 *Nucl. Phys.* **B258** 46–74

[27] Braun V, He Y H, Ovrut B A and Pantev T 2005 *Phys. Lett.* **B618** 252–258 (*Preprint* hep-th/0501070)

[28] Dixon L J, Harvey J A, Vafa C and Witten E 1985 *Nucl. Phys.* **B261** 678–686

[29] Dixon L J, Harvey J A, Vafa C and Witten E 1986 *Nucl. Phys.* **B274** 285–314

[30] Lebedev O *et al.* 2007 *Phys. Lett.* **B645** 88–94 (*Preprint* hep-th/0611095)

[31] Lebedev O, Nilles H P, Ramos-Sánchez S, Ratz M and Vaudrevange P K S 2008 *Phys. Lett.* **B668** 331–335 (*Preprint* 0807.4384)

[32] Lebedev O *et al.* 2008 *Phys. Rev.* **D77** 046013 (*Preprint* arXiv:0708.2691 [hep-th])

[33] Lebedev O *et al.* 2007 *Phys. Rev. Lett.* **98** 181602 (*Preprint* hep-th/0611203)

[34] Buchmüller W, Hamaguchi K, Lebedev O, Ramos-Sánchez S and Ratz M 2007 *Phys. Rev. Lett.* **99** 021601 (*Preprint* hep-ph/0703078)

[35] Uranga A 2010 *J. Phys. Conf. Ser.* **259** 012014

[36] Cremades D, Ibanez L E and Marchesano F 2002 *Nucl. Phys.* **B643** 93–130 (*Preprint* hep-th/0205074)

[37] Gmeiner F and Honecker G 2008 *JHEP* **07** 052 (*Preprint* 0806.3039)

Estados coherentes para potenciales generales

José Récamier
Instituto de Ciencias Físicas
Universidad Nacional Autónoma de México

Resumen

En este curso de presenta la construcción de los estados coherentes de oscilador armónico a partir de sus tres definiciones alternativas [1] y su generalización a álgebras distintas a las del oscilador armónico utilizando el formalismo de operadores deformados [2]. En particular se considera el caso de un potencial tipo Pöschl-Teller trigonométrico con un número infinito de estados ligados y uno modificado con un número finito de estados ligados[3].

1. Introducción

La idea de construir estados cuánticos cuyo comportamiento fuera lo más parecido posible al de los estados clásicos fué propuesta por Schrödinger en los inicios de la mecánica cuántica con el fin de tener un mejor entendimiento del comportamiento de los estados cuánticos. Estos estados fueron concebidos con la condición de que las relaciones de incertidumbre de Heisenbreg se satisfacen como una igualdad, esto es: $\Delta x \Delta p = \hbar/2$ siendo x la coordenada y p el momento conjugado. Cerca de 40 años más tarde, Glauber revivió la idea de estos estados en el estudio de la función de correlación del campo electromagnético y mostró que pueden obtenerse a partir de cualquiera de las siguientes definiciones: *i)* como aquellos estados que se obtienen tras la aplicación del operador de desplazamiento $D(\alpha) = \exp(\alpha\hat{a}^\dagger - \alpha^*\hat{a})$ sobre el estado de vacío del oscilador armónico $D(\alpha)|0\rangle = |\alpha\rangle$, *ii)* como estados propios del operador de aniquilación del oscilador armónico $\hat{a}|\alpha\rangle = \alpha|\alpha\rangle$ y *iii)* como aquellos estados para los cuales se tiene una relación de mínima incertidumbre con $\Delta x = \Delta p$.

Consideremos la primera definición: $D(\alpha)|0\rangle = |\alpha\rangle$. Debido a que el conmutador de los operadores $\hat{a}$, $\hat{a}^\dagger$ es un escalar $[\hat{a}, \hat{a}^\dagger] = 1$, es posible escribir el

operador de desplazamiento en forma de un producto de exponenciales [4] esto es:

$$D(\alpha) = \exp(\alpha\hat{a}^\dagger - \alpha^*\hat{a}) = \exp(-\frac{|\alpha|^2}{2})\exp(\alpha\hat{a}^\dagger)\exp(-\alpha\hat{a}). \tag{1}$$

Utilizando ahora el desarrollo de la exponencial

$$\exp(\alpha\hat{O}) = \sum_{n=0}^{\infty}\frac{\alpha^n}{n!}\hat{O}^n$$

y el hecho de que los operadores $\hat{a}$, $\hat{a}^\dagger$, $\hat{n} = \hat{a}^\dagger\hat{a}$ actuando sobre los estados propios del oscilador armónico (o estados de Fock) $|n\rangle$ dan: $\hat{a}|n\rangle = \sqrt{n}|n-1\rangle$, $\hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle$, $\hat{n}|n\rangle = n|n\rangle$ obtenemos al aplicar la ecuación (1) sobre el estado de vacío $|0\rangle$:

$$|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}}\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}|n\rangle. \tag{2}$$

Demostraremos ahora que los estados dados por la ecuación (2) cumplen con la segunda definición (ser estados propios del operador de aniquilación del oscilador armónico).

$$\begin{aligned}\hat{a}|\alpha\rangle &= e^{-\frac{|\alpha|^2}{2}}\textstyle\sum_{n=0}^{\infty}\frac{\alpha^n}{\sqrt{n!}}\hat{a}|n\rangle = e^{-\frac{|\alpha|^2}{2}}\sum_{n=1}^{\infty}\frac{\alpha\alpha^{n-1}}{\sqrt{(n-1)!}}|n-1\rangle \\ &= \alpha e^{-\frac{|\alpha|^2}{2}}\textstyle\sum_{n=1}^{\infty}\frac{\alpha^{n-1}}{\sqrt{(n-1)!}}|n-1\rangle = \alpha|\alpha\rangle. \end{aligned}\tag{3}$$

Nótese que para llegar a este resultado en el último paso de la demostración utilizamos el hecho de que el oscilador armónico soporta un número infinito de estados ligados.
Finalmente mostraremos que estos estados son estados de mínima dispersión. Para ello recordemos que la dispersión en la observable A se define como:

$$\Delta A = \sqrt{\langle A^2\rangle - \langle A\rangle^2} \tag{4}$$

escribiendo a la coordenada y al momento en términos de los operadores de creación y aniquilación [5]

$$\hat{x} = \sqrt{\frac{\hbar}{2m\omega}}\left(\hat{a} + \hat{a}^\dagger\right)$$
$$\hat{p} = i\sqrt{\frac{\hbar m\omega}{2}}\left(\hat{a}^\dagger - \hat{a}\right)$$

y usando que $\hat{a}|\alpha\rangle = \alpha|\alpha\rangle$ y $\langle\alpha|\hat{a}^\dagger = \langle\alpha|\alpha^*$ obtenemos:

$$\langle\alpha|\hat{x}|\alpha\rangle = \sqrt{\frac{\hbar}{2m\omega}}\,(\alpha + \alpha^*) \tag{5}$$

$$\langle\alpha|\hat{p}|\alpha\rangle = i\sqrt{\frac{\hbar m\omega}{2}}\,(\alpha^* - \alpha) \tag{6}$$

mientras que

$$\langle\alpha|\hat{x}^2|\alpha\rangle = \frac{\hbar}{2m\omega}\left(\alpha^2 + \alpha^{*2} + 2|\alpha|^2 + 1\right) \tag{7}$$

y

$$\langle\alpha|\hat{p}^2|\alpha\rangle = -\frac{\hbar m\omega}{2}\left(\alpha^2 + \alpha^{*2} - 2|\alpha|^2 - 1\right) \tag{8}$$

de donde obtenemos las dispersiones

$$(\Delta\hat{x})^2 = \left(\frac{\hbar}{2m\omega}\right)\left[\alpha^2 + \alpha^{*2} + 2|\alpha|^2 + 1 - (\alpha^2 + 2|\alpha|^2 + \alpha^{*2})\right] = \left(\frac{\hbar}{2m\omega}\right)$$

$$(\Delta\hat{p})^2 = \left(\frac{\hbar m\omega}{2}\right)\left[-\alpha^2 - \alpha^{*2} + 2|\alpha|^2 + 1 + (\alpha^2 - 2|\alpha|^2 + \alpha^{*2})\right] = \left(\frac{\hbar m\omega}{2}\right)$$

y el producto de las incertidumbres queda:

$$\Delta\hat{x}\Delta\hat{p} = \frac{\hbar}{2} \tag{9}$$

independiente del estado coherente $|\alpha\rangle$ de que se trate.

2. Osciladores deformados

Supongamos que tenemos un sistema no perturbado cuyo Hamiltonano podemos escribir en términos de operadores deformados $\hat{A} = \hat{a}f(\hat{n})$, $\hat{A}^\dagger = f(\hat{n})\hat{a}^\dagger$ siendo $f(\hat{n})$ una función de deformación que depende del operador de número y cuya forma definiremos posteriormente. Si consideramos un Hamiltoniano con la estructura correspondiente a la del oscilador armónico y lo expresamos en términos de los operadores deformados obtenemos

$$H_D = \frac{\hbar\Omega}{2}\left(\hat{A}^\dagger\hat{A} + \hat{A}\hat{A}^\dagger\right). \tag{10}$$

A partir de las definiciones de los operadores deformados obtenemos las reglas de conmutación:

$$[\hat{A}, \hat{A}^\dagger] = (\hat{n}+1)f^2(\hat{n}+1) - \hat{n}f^2(\hat{n}), \quad [\hat{A}, \hat{n}] = \hat{A}, \quad [\hat{A}^\dagger, \hat{n}] = -\hat{A}^\dagger \tag{11}$$

de donde el Hamiltoniano deformado H_D puede escribirse de la siguiente manera:

$$H_D = \frac{\hbar\Omega}{2}\left(\hat{n}f^2(\hat{n}) + (\hat{n}+1)f^2(\hat{n}+1)\right) \tag{12}$$

nótese que el límite armónico se obtiene haciendo que la función de deformación tienda a la unidad.

2.1. Potencial de Pöschl-Teller modificado

Este potencial ha sido ampliamente utilizado en diversas áreas de la física, en particular en física atómica y molecular y sus soluciones son conocidas [6] está dado por la expresión

$$V(x) = U_0 \tanh^2(ax) \tag{13}$$

en donde U_0 es la profundidad del pozo, a es el rango del potencial y x es la distancia con respecto a la posición de equilibrio. Las soluciones de la ecuación de Schrödinger y sus valores propios son:

$$\psi_n^\epsilon(\zeta) = N_n^\epsilon \left(1-\zeta^2\right)^{\epsilon/2} F(-n, \epsilon+s+1; \epsilon+1, (1-\zeta)/2) \tag{14}$$

$$E_n = U_0 - \frac{\hbar^2 a^2}{2m}(s-n)^2 = \frac{\hbar^2 a^2}{2m}\left(s + 2ns - n^2\right) \tag{15}$$

en donde N_n^ϵ es una constante de normalización, $\zeta = \tanh(ax)$, m es la masa reducida de la molécula, s está relacionada con la profundidad del potencial de forma tal que $s(s+1) = 2mU_0/\hbar^2a^2$ siendo $\epsilon = \sqrt{-2m(E-U_0)}/\hbar a$ y $F(a,b;c,z)$ es la función hipergeométrica[7]. El número de estados ligados que soporta el potencial está dado por el límite de disociación $\epsilon = s - n = 0$. Eligiendo una función de deformación

$$f^2(\hat{n}) = \frac{\hbar a^2}{2m\Omega}(2s+1-\hat{n}) \tag{16}$$

y substituyendo en la ecuación para el Hamiltoniano deformado (Ec.(12)) obtenemos

$$H_D = \frac{\hbar^2 a^2}{2m}\left(-\hat{n}^2 + 2s\hat{n} + s\right) \tag{17}$$

cuyo espectro es idéntico al dado en la ecuación (15). El conmutador entre los operadores deformados queda:

$$[\hat{A}, \hat{A}^{\dagger}] = \frac{\hbar a^2}{m\Omega} s(1 - \frac{\hat{n}}{s}) \equiv \chi \left(1 - \frac{\hat{n}}{s}\right). \tag{18}$$

En términos del parámetro de anarmonicidad χ la función de deformación queda

$$f^2(\hat{n}) = \chi \left(1 - \frac{\hat{n} - 1}{2s}\right). \tag{19}$$

2.2. Potencial de Pöschl-Teller trigonométrico

Este potencial tiene la forma

$$V(x) = U_0 \tan^2(ax) \tag{20}$$

siendo U_0 la profundidad del pozo, a su rango y x la distancia a la posición de equilibrio. Las funciones propias y valores propios para este potencial son[8]:

$$\Psi_n(x) = \sqrt{\frac{a(\lambda + n)(\Gamma(2\lambda + n)}{\Gamma(n + 1)}} (\cos(ax))^{1/2} P^{1/2-\lambda}_{n+\lambda-1/2}(\sin(ax)) \tag{21}$$

$$E_n = \frac{\hbar^2 a^2}{2\mu}(n^2 + 2\lambda n + \lambda) \tag{22}$$

en donde μ es la masa de la partícula y el parámetro λ está relacionado con la intensidad del potencial y su alcance mediante la expresión $\lambda(\lambda + 1) = 2\mu U_0/\hbar^2 a^2$. Nótese que el número de excitaciones n puede tomar cualquier valor mayor o igual a cero, esto es, se tiene un caso similar al del oscilador armónico en donde el potencial soporta un número infinito de estados ligados. Si proponemos la función de deformación

$$f^2(\hat{n}) = \frac{\hbar a^2}{2\mu\Omega}(\hat{n} + 2\lambda - 1) \tag{23}$$

el Hamiltoniano deformado H_D queda:

$$H_D = \frac{\hbar^2 a^2}{2\mu}(\hat{n}^2 + 2\lambda\hat{n} + \lambda) \tag{24}$$

cuyo espectro es idéntico con el dado en la ecuación 22. Las relaciones de conmutación entre los operadores deformados son:

$$[\hat{A}, \hat{A}^\dagger] = \frac{\hbar a^2}{\mu\Omega}(\lambda + \hat{n}) \equiv \chi' \left(1 + \frac{\hat{n}}{\lambda}\right) \tag{25}$$

en donde $\chi' = \hbar a^2 \lambda / \mu\Omega$.

3. Estados coherentes no lineales

Para sistemas que pueden ser descritos por Hamiltonianos que contienen términos de orden superior al lineal en el operador de número, la construcción de sus estados coherentes depende de la definición que se trate de generalizar a diferencia del caso armónico en donde los estados coherentes que se obtienen son independientes de la definición utilizada[9].

3.1. Estados propios del operador de aniquilación deformado

Una vez elegida la función de deformación, los operadores $\hat{A}$, $\hat{A}^\dagger$ quedan especificados. Una posible definición de los estados coherentes consiste en aquellos estados que son estados propios del operador de aniquilación deformado, esto es:

$$\hat{A}|\alpha, f\rangle = \alpha|\alpha, f\rangle \tag{26}$$

puede mostrarse fácilmente que los estados $|\alpha, f\rangle$ propuestos en la referencia [2]

$$|\alpha, f\rangle = N_{\alpha,f} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!} f(n)!} |n\rangle \tag{27}$$

en donde $f(n)! = f(1)f(2)\cdots f(n)$ y $N_{\alpha,f}$ es una constante de normalización son solución de la ecuación (26). Nótese que la suma corre hasta infinito para que la expresión dada para el estado coherente sea solución.

3.1.1. Estados coherentes para el potencial de Pöschl-Teller modificado

En este caso se cuenta con un potencial que soporta s-1 estados ligados, esto hace que la expresión que encontramos arriba (ver ecuación 27) sea

solamente aproximada.

$$|\alpha, f\rangle = N_{\alpha,f} \sum_{n=0}^{s-1} \frac{\alpha^n}{\sqrt{n!} f(n)!} |n\rangle \tag{28}$$

con la constante de normalización

$$N_{\alpha,f} = \left(\sum_{n=0}^{s-1} \frac{|\alpha|^{2n}}{n![f(n)!]^2} \right)^{-1/2}.$$

De hecho, la ecuación de eigenvalores da:

$$\hat{A}|\alpha, f\rangle = \alpha|\alpha, f\rangle - \frac{N_{\alpha,f}\alpha^s}{\sqrt{(s-1)!} f(s-1)!} |s-1\rangle \tag{29}$$

la importancia del segundo término del lado derecho de la ecuación anterior depende del número de estados ligados que soporta el potencial así como del *tamaño* del estado coherente. Para potenciales que soportan un número grande de estados ligados es posible alcanzar valores razonablemente grandes de $|\alpha|$ y contar con una buena aproximación. Si el número de estados ligados es pequeño, entonces debemos restringirnos a valores pequeños de $|\alpha|$. Sustituyendo la forma explícita de la función de deformación obtenemos:

$$|\alpha, f\rangle = N_{\alpha,f} \sum_{n=0}^{s-1} \sqrt{\frac{\Gamma(2s+1)}{\Gamma(2s+1-n)}} \left(\frac{2s}{\chi} \right)^{n/2} \alpha^n |n\rangle \tag{30}$$

en donde $\chi = \hbar a^2 s / m\Omega$.

3.1.2. Estados coherentes para el potencial de Pöschl-Teller trigonométrico

En este caso se cuenta con un potencial que soporta un número infinito de estados ligados por lo cual la expresión dada en la ecuación 27 es exacta. Sustituyendo la forma explícita de la función de deformación obtenemos

$$|\alpha, f\rangle = N_{\alpha,f} \sum_{n=0}^{\infty} \sqrt{\frac{\Gamma(2\lambda)}{n!\Gamma(2\lambda+n)}} \left(\frac{2\mu\Omega}{\hbar a^2} \right)^{n/2} \alpha^n |n\rangle \tag{31}$$

siendo $N_{\alpha,f}$ la constante de normalización

$$N_{\alpha,f} = \left(\sum_{n=0}^{\infty} \frac{|\alpha|^{2n}}{n![f(n)!]^2} \right)^{-1/2}.$$

Es posible mostrar que estos estados forman un conjunto completo, esto es, una base.

3.2. Operador de desplazamiento generalizado

Escribiendo el operador de desplazamiento $D(\alpha)$ en términos de los operadores deformados $\hat{A}$, $\hat{A}^\dagger$ se obtiene

$$D_D(\alpha) = \exp\left(\alpha \hat{A}^\dagger - \alpha^* \hat{A}\right) \tag{32}$$

el problema que se tiene ahora es que el conmutador entre los operadores deformados no es un escalar (ver las ecuaciones 11), sino una función del operador de número y eso hace que el conjunto de operadores formado por los operadores deformados y el resultado de su conmutador no forme, en general, un álgebra de Lie finita. Debido a este hecho, no es posible expresar la exponencial de la suma (el operador de desplazamiento deformado) en términos de un producto de exponenciales. Sin embargo, en aquellos casos en que el conmutador es una función lineal del operador de número sí es posible hacerlo ya que el conjunto de operadores $\{\hat{A}, \hat{A}^\dagger, \hat{n}, 1\}$ forma un álgebra de Lie finita.

Los casos que hemos considerado aquí cumplen con esta característica (ver las ecuaciones 18 y 25). El estado coherente que se obtiene al aplicar el operado de desplazamiento deformado al estado de vacío resulta ser [3, 10]:

$$D_D(\alpha)|0\rangle = N_D(\alpha) \sum_{k=0}^{\infty} \frac{\gamma^k}{\sqrt{k!}} f(k)!|k\rangle \tag{33}$$

en donde $N_D(\alpha)$ es una constante de normalización.

3.2.1. Estados coherentes para el potencial de Pöschl-Teller modificado

En este caso las relaciones de conmutación son:

$$[\hat{A}, \hat{A}^\dagger] = \frac{\hbar a^2}{m\Omega}(s - \hat{n}), \quad [\hat{A}, \hat{n}] = \hat{A}, \quad [\hat{A}^\dagger, \hat{n}] = -\hat{A}^\dagger$$

por lo tanto el operador de desplazamiento puede escribirse en forma de un producto de exponenciales. El resultado es [11]

$$D_D(\alpha) = \exp\left[\alpha\frac{\tan\left(|\alpha|\sqrt{\chi/2s}\right)}{|\alpha|\sqrt{\chi/2s}}\hat{A}^\dagger\right]\exp\left[\frac{ln\left(\cos(|\alpha|\sqrt{\chi/2s}\right)}{\chi/2s}\chi(1-\frac{\hat{n}}{s})\right] \times \exp\left[-\alpha^*\frac{\tan\left(|\alpha|\sqrt{\chi/2s}\right)}{|\alpha|\sqrt{\chi/2s}}\hat{A}\right] \quad (34)$$

aplicando este operador al estado de vacío obtenemos:

$$|\zeta\rangle = D_D(\zeta)|0\rangle = \frac{1}{(1+|\zeta|^2)^s}\sum_n \frac{\zeta^n}{\sqrt{(\chi/2s)^n}}\frac{f(n)!}{\sqrt{n!}}|n\rangle \quad (35)$$

Teniendo en cuenta que este potencial soporta s-1 estados ligados, obtenemos los estados coherentes aproximados

$$|\zeta\rangle \simeq \frac{1}{(1+|\zeta|^2)}\sum_{n=0}^{s-1}\sqrt{\frac{\Gamma(2s+1)}{n!\Gamma(2s+1-n)}}\zeta^n|n\rangle. \quad (36)$$

3.2.2. Estados coherentes para el potencial de Pöschl-Teller trigonométrico

En este caso las relaciones de conmutación entre los operadores deformados son

$$[\hat{A},\hat{A}^\dagger] = \frac{\hbar a^2}{\mu\Omega}(\lambda+\hat{n}), \quad [\hat{A},\hat{n}] = \hat{A}, \quad [\hat{A}^\dagger,\hat{n}] = -\hat{A}^\dagger$$

el operador de desplazamiento puede escribirse en forma de un producto de exponenciales y dado que este potencial soporta un número infinito de estados ligados (esto debido a la diferencia en el signo del lado derecho de los conmutadores entre los operadores deformados), las funciones trigonométricas son reemplazadas por funciones hiperbólicas. En este caso se tiene una expresión exacta para los estados coherentes la cual se obtiene al aplicar el operador de desplazamiento deformado al estado de vacío. El resultado es:

$$|\zeta,f\rangle = (1-|\zeta|^2)^\lambda\sum_{n=0}^{\infty}\sqrt{\frac{\Gamma(2\lambda+n)}{n!\Gamma(2\lambda)}}\zeta^n|n\rangle. \quad (37)$$

Referencias

[1] R J Glauber, Phys. Rev. Lett. **10**, 84 (1963).

[2] V I Man'ko, G Marmo, F Zaccaria and ECG Sudarshan, Phys. Scr. **55**, 528 (1997).

[3] O. de los Santos and J. Récamier, J. Phys. A: Math. Theor. **44**, (2011) 145307

[4] Merzbacher E 1970 *Quantum mechanics*, (Second edition, Wiley, New York)

[5] de la Peña L 1991 *Introducción a la mecánica cuántica*, Fondo de Cultura Económica, México.

[6] Landau L and Lifshitz E 1967 *Mechanique Quantique: Theórie non Relativiste* (Moscou: Deuxiéme édition MIR) p 94.

[7] M. Abramowitz and I A Stegun, 1972 *Handbook of Mathematical Functions* (New York, Dover) p 555

[8] M M Nieto, Phys. Rev. A **17**, 1273 (1978)

[9] W M Zhang W, D H Feng and R Gilmore, Reviews of Modern Physics **62**, 868 (1990).

[10] R. Román-Ancheyta, O. de los Santos-Sánchez and J. Récamier, J. Phys A: Math. Theor. **44**, (2011) 435304

[11] *Estados coherentes no lineales para potenciales generales*, Tesis Doctoral, Octavio de los Santos Sánchez, BUAP, junio 2011.

De la electrónica a la fotónica

Jazael Gómez, Julia Tagüeña-Martínez, Rocío Nava y Jesús Antonio del Río
Centro de Investigación en Energía,
Universidad Nacional Autónoma de México,
A.P. 34, 62580 Temixco, Morelos, México.

"Todo áquel que no queda fuertemente impresionado por la teoría cuántica es porque no la ha entendido"
Niels Bohr (1885-1962)

Resumen

Empezamos por dar un breve repaso sobre algunos conceptos de mecánica cuántica relacionados con la luz, los electrones y los materiales llamados semiconductores, en particular sobre el silicio. Resaltamos el papel fundamental del silicio en el desarrollo de la electrónica. Explicamos cómo el silicio puede hacerse poroso y nanoestructurado, con lo que se puede utilizar para fabricar dispositivos fotónicos. La fotónica está causando en el siglo XXI una revolución equivalente a la que causó la electrónica en el siglo XX. Como ejemplo damos la construcción de un filtro que simula la acción del agua marina al paso de la luz, realizado en el Centro de Investigación en Energía de la UNAM, como parte de una tesis profesional.

Introducción

El propósito de estas notas es dar un ejemplo de cómo el silicio, material fundamental en electrónica, puede también ser aprovechado en aplicaciones fotónicas. Este capítulo es una actualización del texto presentado en el curso de verano de 2009[1].

Para entrar en materia decidimos revisar algunos conceptos básicos de la mecánica cuántica sobre la luz y los electrones[2], en un recorrido histórico de cómo se llegó a sentar las bases de la visión moderna de la materia. En particular resaltamos el comportamiento de los materiales semiconductores.[3,4]

Si bien la existencia de los átomos es ya un conocimiento asimilado por la sociedad, los efectos cuánticos no son intuitivos ni suelen formar parte de la cultura científica, a pesar de que la tecnología que define nuestra forma de vida se basa en ellos.

Desde el invento del primer transistor, predicho por la mecánica cuántica, en 1948, la electrónica se ha basado sobre todo en un material semiconductor: el silicio. Antes de esto se usaban los tubos evacuados o bulbos, que requerían de equipos de gran tamaño; sin embargo, con el transistor el tamaño de los dispositivos electrónicos se ha ido reduciendo cada vez más. En la década de los setenta se desarrolló el circuito integrado, que puede contener centenares de transistores y permite construir los chips de las computadoras y los satélites de comunicaciones. Los circuitos integrados revolucionaron los campos de la comunicación y de la informática y a partir de ellos se creó una poderosa industria. También el silicio es componente fundamental de aproximadamente el 95% de las celdas solares, construidas gracias a la investigación espacial. Sin embargo, ese camino alcanzó un límite y ahora se investigan nuevas tecnologías, con diferentes principios de operación y diferentes materiales.

Hablaremos aquí de otro posible camino tecnológico: la nanotecnología. El propio silicio puede hacerse poroso[5,6] (SP) y se vuelve un material nanoestructurado con el que se pueden construir dispositivos llamados "fotónicos"[1]. La fotónica en su acepción más general se refiere a fenómenos y aplicaciones en que la luz (compuesta de fotones) se usa para procesar o transmitir información o para modificar materiales[7,8]. Posiblemente el mejor ejemplo actual de la fotónica es la fibra óptica, que también está hecha de silicio. En el Centro de Investigación en Energía (UNAM) se están estudiando multicapas de silicio poroso tanto experimental como teóricamente que son estructuras fotónicas que pueden filtrar o reflejar la luz. Como ejemplo presentamos un filtro diseñado en una tesis profesional[9], que simula el efecto que sufre la luz cuando atraviesa el agua del mar, afectando así a los seres vivos que la reciben.
Una vez más, a través de la nanotecnología y la fotónica, el silicio seguirá siendo clave en el siglo XXI.

Sobre la luz ¿onda o partícula?
La mecánica newtoniana fue y sigue siendo tan exitosa, que no es sorprendente que Isaac Newton (1646-1727) describiera a la luz como formada por partículas diminutas, por corpúsculos. Con este modelo era fácil explicar el movimiento rectilíneo de la luz los experimentos de reflexión y refracción. Sin embargo, su contemporáneo, el físico holandés Christian Huygens (1629-1665) desarrolló una teoría en donde la luz era una onda, como las que vemos en el agua, que además de poder explicar la reflexión y la refracción, además podía explicar los patrones de interferencia y su comportamiento cuando choca con un obstáculo, como lo probó años después el inglés Thomas Young (1773-1829) a principios del siglo XIX y poco después el francés Agustín Fresnel (1788-1827). Cierto que al hacer una analogía entre la luz y las ondas en el agua, hubo que inventar un medio invisible, al que se llamó éter, por el que se propagaba la luz.
Los experimentos de Young fueron considerados en Gran Bretaña como antipatrióticos por el enorme prestigio que tenía Newton. Sin embargo, cuando León Foucault (1819-1868) probó que la velocidad de la luz es menor en el agua que en el aire, todo el mundo aceptó que la luz era una onda cuya velocidad dependía del medio que se moviera. Cuando el gran físico escocés James Clerk Maxwell (1831-1879) probó la existencia de las ondas electromagnéticas, producto de cambios en los campos eléctricos y magnéticos y en 1887 Heinrich Hertz (1857-1894) logró transmitir radiación electromagnética que hoy llamamos ondas de radio, parecía que la teoría ondulatoria de la luz era ya indiscutible. Pero la física moderna tenía reservadas unas cuantas sorpresas que llevaron a recuperar la teoría corpuscular.

Sobre los átomos y sus partículas
Todos hemos escuchado la historia de que los griegos clásicos imaginaron la existencia de átomos. Demócrito (460 AC- 370 AC) escribió "Las únicas realidades existentes son los átomos y el espacio vació; lo demás es mera especulación". Sin embargo, la teoría aristotélica de los cuatro elementos,

agua, tierra, fuego y aire, fue la que prevaleció aproximadamente 2000 años. Aunque varios científicos retomaron la teoría atómica, fue el químico francés Antoine Lavosier (1743-1794) el que identificó muchos elementos puros y probó que en la combustión el oxígeno del aire se combina con otros elementos. El estudio de los gases a mediados del siglo XIX contribuyó enormemente a la hipótesis atómica pues con ella se podían explicar sus propiedades, imaginando al gas como compuesto de muchas partículas que obedecen las leyes de la mecánica. Al ser tantas partículas en número, estos estudios llevaron a lo que hoy llamamos mecánica estadística.
A pesar de estos éxitos, la primera prueba directa de su existencia es bastante reciente y se debe, como tantos otros logros de la física moderna, a Albert Einstein (1879-1955), cuando decidió explicar el llamado movimiento browniano.
El botánico inglés Thomas Brown (1773-1858) había reportado que al observar al microscopio polen flotando sobre agua se observa que el polen sigue un movimiento aleatorio que se llama movimiento browniano. Einstein probó que este movimiento sigue una ley estadística producto de los golpes de los átomos en movimiento, sobre los granos de polen.

Los electrones

A finales del siglo XIX se conocían los llamados rayos catódicos que era una radiación a través de un tubo al vacío, producida cuando por un hilo metálico circula una corriente. Contribuyó a la confusión el descubrimiento de los Rayos X por Wilhelm Röntgen (1845-1923) en 1895. El inglés J.J. Thomson trabajaba en esa época en Cambridge (1856-1940) y diseñó un experimento para estudiar partículas cargadas en movimiento, al desviar su trayectoria con campos eléctricos y magnéticos. Así descubrió que los rayos catódicos eran las partículas que llamamos electrones. Encontró la relación entre la carga y la masa del electrón y que este resultado era el mismo cualquiera que fuera el metal utilizado, lo que le llevó a concluir que los electrones eran parte fundamental del átomo. Como los átomos son neutros tenía que existir una contraparte positiva. El propio Thomson estudió otros rayos positivos, que resultaron iones de diferentes elementos. Sin embargo, fue Ernest Rutherford (1871-1937) quien propuso el modelo del átomo que es la base de lo que sabemos hoy: un núcleo positivo y una nube de electrones que rodea al núcleo. El por qué los electrones negativos no colapsan en el núcleo positivo se logró entender justamente al estudiar la interacción de los electrones con la luz, que es uno de los primeros aciertos de la mecánica cuántica.

Los cuantos de luz

A un cuerpo que es un perfecto absorbedor o emisor de radiación se le llama “cuerpo negro”. Es un nombre que confunde un poco porque no tiene que ser en verdad negro. Por ejemplo, un cuerpo negro puede ser una caja a la que se le haga un agujero por el que penetra radiación luminosa. Esta radiación rebotará en las paredes, será absorbida y es muy poco probable que vuelva a salir. La teoría clásica para estudiar la radiación del cuerpo negro se ajustaba bien a la zona de las bajas frecuencias, pero predecía incorrectamente que debería de haber grandes cantidades de energía en la zona del ultravioleta, lo

que se llamó la catástrofe del ultravioleta. Max Planck (1858-1947) logró reproducir la curva experimental del espectro del cuerpo negro; su fórmula que fue hecha pública en 1900 introducía una hipótesis revolucionaria: la luz se comporta como partículas a las que llamó cuantos. Cada quantum tiene una energía E igual al producto de la constante de Planck h ($h = 6.63 \times 10^{-34}$ joules) por su frecuencia ν, $E = h\nu$. Planck argumentó que los osciladores interiores de los átomos estaban cuantizados y que sólo podían emitir energía en ciertos paquetes.

Es Albert Einstein el que utilizando la propuesta de Planck lleva a su verdadero impacto cuando logra explicar el efecto fotoeléctrico, descubierto por Lenard (1862-1947) en 1899. En sus experimentos Lenard demostró que cuando la luz incide sobre un metal situado al vacío, éste emite electrones. Observó, al utilizar luz monocromática, que aunque cambiara la intensidad de la luz, los electrones salían con la misma velocidad, a menos que cambiara la frecuencia de la luz incidente. Lo que probó Einstein es que como la luz se puede comportar como cuantos de energía, éstos chocan con los electrones y le proporciona esta misma cantidad de energía $h\nu$. Si hay un cambio de color en la luz, entonces sí cambiará la energía comunicada. Por este trabajo Einstein recibió el premio Nobel en 1921. La medición precisa de la constante de Planck y la relación precisa de la longitud de onda de la luz y la velocidad de los electrones tardó unos 10 años y la realizó el físico estadounidense Robert Millikan (1868-1953), quien también recibió el premio Nobel en 1923.

Con estas ideas de cuantización el famoso físico danés Niels Bohr (1885-1962) hizo su famoso modelo, afirmando que los electrones en los átomos estaban en niveles fijos de energía y sólo podían cambiar de órbita con emisión de radiaciones múltiplos de un cuanto elemental. Su modelo sirvió para explicar el espectro de luz del átomo de hidrógeno, pero falla en los demás elementos y ya ha sido ampliamente superado. Sin embargo, continúa siendo muy popular por su similitud al sistema planetario.

Fotones y electrones

El nombre de fotón para el cuanto o partícula de luz no se adoptó hasta 1926 y fue introducido por el estadounidense Gilbert Lewis (1875-1946) y se volvió popular después del quinto congreso de Solvay en 1927 que se llamó "Electrones y fotones". Se contaba entonces con dos teorías de la luz, la ondulatoria y la corpuscular. Ambas eran imprescindibles y llevaban a lo que se volvió la llamada dualidad onda partícula. El paso siguiente a la dualidad en la luz fue la dualidad de la materia.

Un noble francés, Louis de Broglie (1892-1987) sugirió que si la luz que es una onda se comporta como una partícula ¿por qué no se podría comportar un electrón como onda? Por el desarrollo matemático de esta idea durante su doctorado recibió el premio Nobel en 1929. Utilizó las ecuaciones de Plank y de Einstein para los cuantos de luz $E = h\nu$ y $p = h\nu/c$ donde p es el momento y c la velocidad de la luz en el vacío, donde se mezclan las propiedades ondulatorias y corpusculares y se aplican tanto a la luz como a los electrones. La prueba del comportamiento ondulatorio de los electrones la encontraron dos científicos estadounidenses Clinton Davisson (1881-1958) y Lester Germer (1896-1971) que probaron que los electrones se difractaban por una red cristalina, como si fueran ondas. George Thomson (1892-1975), el hijo de J.J.

Thomson, realizó en Inglaterra un experimento diferente que probó lo mismo y Davisson y Thomson compartieron el premio Nobel en 1937. Es interesante que J.J. Thomson obtuvo el premio Nobel por probar que el electrón era una partícula y su hijo por probar que el electrón era una onda. Ambos tenían razón.

La mecánica cuántica y la física del estado sólido

Hemos dado sólo unas pinceladas de mecánica cuántica que describen la ruptura con la física clásica, pero no lo vemos así en nuestra experiencia cotidiana. Para nosotros esta mezcla partícula onda no es evidente: una pelota es siempre una partícula. Es en el mundo de lo muy pequeño donde los aspectos ondulatorios y corpusculares se intercalan y son igualmente importantes. Sin embargo, aunque no nos demos cuenta de los efectos cuánticos, vivimos en una sociedad totalmente definida por la mecánica cuántica gracias a la tecnología.

Con el desarrollo de la mecánica cuántica y las técnicas de difracción de rayos X para estudiar la materia surge una nueva disciplina, la física del estado sólido cuyas aplicaciones, en particular a través del desarrollo de la electrónica, han cambiado nuestras sociedades: los transistores, las computadoras, las celdas solares y muchos dispositivos electrónicos dominan nuestras comunicaciones, la industria, los transportes y nuestra forma de vida. En toda esta revolución tecnológica los materiales semiconductores, entre los que destaca el silicio, son claves por sus propiedades especiales que comentaremos en un momento.

Más recientemente, a mediados del siglo XX, surge la fotónica, que aprovecha a los fotones de manera similar a la que la electrónica lo hace con los electrones. La pregunta que vamos a contestar es si el silicio puede ser también útil en la fotónica.

¿Qué es un semiconductor?

Muchas de las características de los semiconductores se observaron antes de que se entendieran[2,3]. En el siglo XIX el famoso Michael Faraday, pionero del electromagnetismo, notó que ciertos materiales como el sulfuro de plata, aunque sí conducían la electricidad, tenían un comportamiento anómalo con la temperatura. Sus características conductoras mejoraban a medida que aumentaba la temperatura, a diferencia del comportamiento de los metales, que se vuelven peores conductores cuando están más calientes. También se observó que estos materiales eran muy sensibles a la luz y conducían mejor al ser iluminados. Sin embargo, fue sólo hasta el siglo XX, con la mecánica cuántica, que se penetró al mundo de los átomos y las propiedades electrónicas de los materiales.

Es justamente por sus propiedades electrónicas que los sólidos se clasifican en conductores, semiconductores y aislantes. Para conducir la electricidad, un material requiere de electrones casi libres que puedan moverse. Los electrones en un sólido tienen valores de energía y el moverse implica pasar a estados vacíos de energía.

Cuando los átomos se unen para formar sólidos, sus niveles de energía se agrupan en las llamadas bandas de energía. La teoría de bandas es tal vez el logro principal de la física del estado sólido, desarrollada a partir de la

mecánica cuántica. En la figura 1 se puede ver un diagrama esquemático de bandas de energía[3]. En el caso de los conductores los electrones pueden moverse libremente por la banda de conducción, pues tienen estados libres que pueden tomar. En los aislantes y los semiconductores hay una brecha prohibida de energía. En los aislantes los electrones no “saltan” a la banda de conducción y por lo tanto, no conducen electricidad, no tienen estados libres a dónde ir. En cambio, en los semiconductores, algunos electrones sí pueden tener energía suficiente para saltar (cuando se calientan o se iluminan, lo que explica las observaciones mencionadas) y al llegar a la banda de conducción producen una corriente. Lo más interesante, es que los electrones que saltan dejan huecos en la banda de valencia y estos hoyos también pueden conducir[3]. Aunque un hoyo es de hecho la ausencia de un electrón, se comporta como si fuera una partícula de carga positiva.

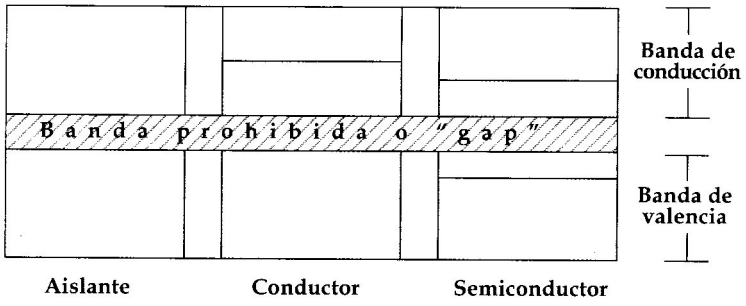


Figura 1. Esquema de diagrama de bandas

La existencia de electrones y hoyos electrónicos es la característica fundamental de los semiconductores. La aplicación de los semiconductores en electrónica se inicia a partir del descubrimiento de que se les puede introducir impurezas que aumenten el número de hoyos (semiconductor tipo p por positivo) o el número de electrones (tipo n, por negativo). Estos semiconductores envenenados o “dopados”, como se llaman, sí alcanzan mayores conductividades y al unirse forman las uniones n-p que tienen propiedades muy importantes: por ejemplo, pueden ser rectificadoras, que dejan pasar la corriente en una dirección y no en la otra. Un transistor, descubierto en 1948, está hecho de combinaciones de uniones p-n-p o n-p-n y tiene diferentes funciones. Puede amplificar señales de corriente o voltaje, puede modular señales o simplemente actuar como interruptor, como en las computadoras. Debido a su tamaño reducido, a su bajo costo y a su enorme eficiencia, su aparición sustituyó casi completamente a la tecnología de bulbos al vacío. En una unión n-p, al ser iluminada, se crean pares electrón-hoyo en la unión y se presenta el efecto fotovoltaico, que permite obtener una corriente y es base de las celdas solares, descubiertas para aplicaciones espaciales en 1954.

El silicio es un elemento químico semiconductor de número atómico 14 y pertenece al grupo IVA. Es el segundo elemento más abundante en la corteza terrestre después del oxígeno. El nombre silicio deriva del latín silex (pedernal).

Las herramientas y las armas hechas de pedernal fueron los primeros utensilios de la humanidad. El vidrio es también un silicato. El silicio, a diferencia del carbono, no existe libre en la naturaleza. Como dióxido se encuentra en varias formas de cuarzo. La arena es en gran parte dióxido de silicio (sílice). Su nombre en inglés, *silicon*, le dio el nombre a la zona industrial más famosa del mundo: el *Silicon Valley* (Valle del Silicio) en California, Estados Unidos. El silicio cristalino (c-Si) tiene la misma estructura que el diamante: una red fcc con una base de dos átomos de silicio en (0,0,0) y (1/4,1/4,1/4). La conductividad eléctrica se c-Si incrementa introduciendo impurezas de elementos del grupo III (B, Al, Ga, In) de la tabla periódica para producir c-Si tipo p, donde los huecos son los portadores mayoritarios, o con elementos del grupo V (P, As, Sb) para producir c-Si tipo n, donde los electrones son los portadores de carga mayoritarios. El uso de los semiconductores, mayoritariamente el silicio, ha permitido hacer componentes electrónicas cada vez más pequeñas, los llamados circuitos integrados. Sus aplicaciones son enormes. Han aumentado la potencialidad de las computadoras con el uso de "chips", donde se colocan un gran número de circuitos electrónicos. También se utiliza en las celdas solares. El desarrollo tecnológico ha ido acompañado de reducción de precios.

Nanotecnología

En la microelectrónica se ha reducido mucho el tamaño, pero un transistor tiene todavía millones de átomos y lo podemos ver con un microscopio. En cambio en la nanotecnología se manejan átomos individuales que no podemos ver. En los microcircuitos las unidades son micrómetros, millonésimas de metro, en cambio la nanotecnología se mide en nanómetros (nm) unidades mil veces más pequeñas.

La tecnología es la actividad humana que genera herramientas o procesos a través de transformar los conocimientos científicos en aplicaciones para el desarrollo económico de una sociedad. Se trata de nanotecnología cuando estas herramientas son del orden de nanómetros, es decir de mil millonésimas partes de un metro. Algunas propiedades físicas de los materiales cambian cuando se llega a esos tamaños. Un átomo mide la décima parte de un nanómetro. Si un átomo fuera del tamaño de una canica, una molécula compleja sería del tamaño de nuestro puño. Un cabello humano tiene 10,000 nm de diámetro. La nanotecnología es en realidad una tecnología donde se manipulan moléculas. Sin duda, la nanotecnología es una de las tecnologías del siglo XXI, de la cual podemos esperar muchas sorpresas. Cuando hablamos de su potencial parece algo de ciencia ficción: nanopartículas que atacan tumores cancerosos, almacenamiento de información en tamaños pequeñísimos, telas inteligentes que pueden detectar si estamos enfermos y otras más. Estamos en el nacimiento de esta revolución tecnológica y es importante que México no salga de esta carrera. Muchos grupos están trabajando tanto en nanociencia (que es sobre todo mecánica cuántica), como nanotecnología.

El propio silicio puede hacerse poroso y se vuelve un material nanoestructurado con el que se pueden construir dispositivos fotónicos,

presentando propiedades ópticas novedosas. En el Centro de Investigación en Energía se están estudiando multicapas de silicio poroso tanto experimental como teóricamente[1].

Como hemos dicho, el c-Si es fundamental en electrónica, pero no se emplea en la optoelectrónica debido a su baja eficiencia como emisor de luz, menor que 0.001%. La fotoluminiscencia en un semiconductor involucra la excitación de un electrón de la banda de valencia a la banda de conducción y la subsiguiente recombinación del electrón con un hueco. El c-Si es un semiconductor de brecha energética indirecta (Eg=1.1 eV), donde el mínimo de la banda de conducción y el máximo de la banda de valencia se encuentran en diferentes valores de momento cristalino. Este proceso es menos probable que la recombinación óptica directa, es por ello que el c-Si presenta luminiscencia extremadamente deficiente y sólo en la región del infrarrojo. En contraste, el silicio poroso (SP) es un material nanoestructurado que presenta fotoluminiscencia a una energía mayor que la brecha energética del c-Si y ésta se puede modular en la región del espectro visible.

El SP se produce mediante el ataque electroquímico de una oblea de c-Si en una solución de ácido fluorhídrico (HF). La reacción electroquímica se inicia por la presencia de huecos en la superficie, ya sea procedentes de las impurezas del c-Si o inducidos por iluminación, que reaccionan con el HF al aplicar una corriente anódica. La estructura del SP varía dependiendo de las condiciones de ataque electroquímico, ésta puede ser similar a la de un coral con un esqueleto cristalino muy fino del orden de unos cuantos nanómetros hasta ser un arreglo ordenado de macroporos. Debido a que la reacción electroquímica depende de la densidad de portadores de carga positivos, la resistividad eléctrica del sustrato de c-Si determina el tipo de estructura del SP. Así por ejemplo, en el SP producido a partir de sustratos de c-Si del tipo p con una resistividad de 1 – 10 Ω–cm, el grosor del esqueleto es del orden de 5-10 nm. Como consecuencia hay efectos de confinamiento cuántico dentro del esqueleto que abren la brecha energética del silicio que da lugar a una emisión luminiscente. En cambio, en el SP producido a partir de sustratos de c-Si tipo p con una resistividad de 10^{-2} a 10^{-3} Ω-cm, el grosor de esqueleto es del orden de 10-20 nm. Éste presenta una emisión luminiscente baja y sólo en el infrarrojo, pero se puede variar la porosidad, y con ello el índice de refracción, en un amplio rango para crear estructuras fotónicas de multicapas.

Multicapas de silicio poroso

La síntesis de multicapas de SP se basa en el hecho de que el ataque electroquímico es auto limitante y sólo ocurre en las puntas de los poros donde hay mayor concentración de huecos. Es decir, la región donde ya se han formado los poros no se afecta por el subsiguiente ataque electroquímico. Al alternar la intensidad de la densidad de corriente durante el ataque electroquímico se produce un perfil de índices de refracción de acuerdo la secuencia de capas requerida para crear estructuras fotónicas.

El arreglo experimental para producir multicapas de SP se muestra en figura 2. Éste consta básicamente de una celda electroquímica, una bomba de circulación para el electrolito, una fuente de corriente y una computadora. La celda se conforma por un recipiente de teflón en el que se aloja un

condensador de placas paralelas, donde el cátodo es una malla de platino y el ánodo la oblea de c-Si. Una cara de la oblea de c-Si queda expuesta al electrolito y en la otra se deposita una película de aluminio para hacer contacto eléctrico. Normalmente se sintetizan capas delgadas un espesor del orden de 100 nm, por lo que el tiempo de ataque es de unos cuantos segundos y se requiere del control preciso mediante una computadora.
En la producción de las multicapas fotónicas de SP empleamos sustratos de c-Si con una resistividad eléctrica del orden 10^{-3} Ω-cm tipo p dopado con boro y con una orientación (100), siendo ésta la dirección preferencial de ataque del c-Si. El electrolito se compone de una solución acuosa de Etanol, HF y glicerina Las densidades de corriente aplicadas son J_A=5mA/cm^2 para la capa de baja porosidad y J_B=45mA/cm^2 para la capa de alta porosidad, que corresponden a índices de refracción de 2.1 y 1.4, respectivamente. Alternando la densidad de corriente aplicada durante el ataque electroquímico se pueden producir diversas estructuras, donde el arreglo de capas y los espesores varían de acuerdo a la aplicación en particular.

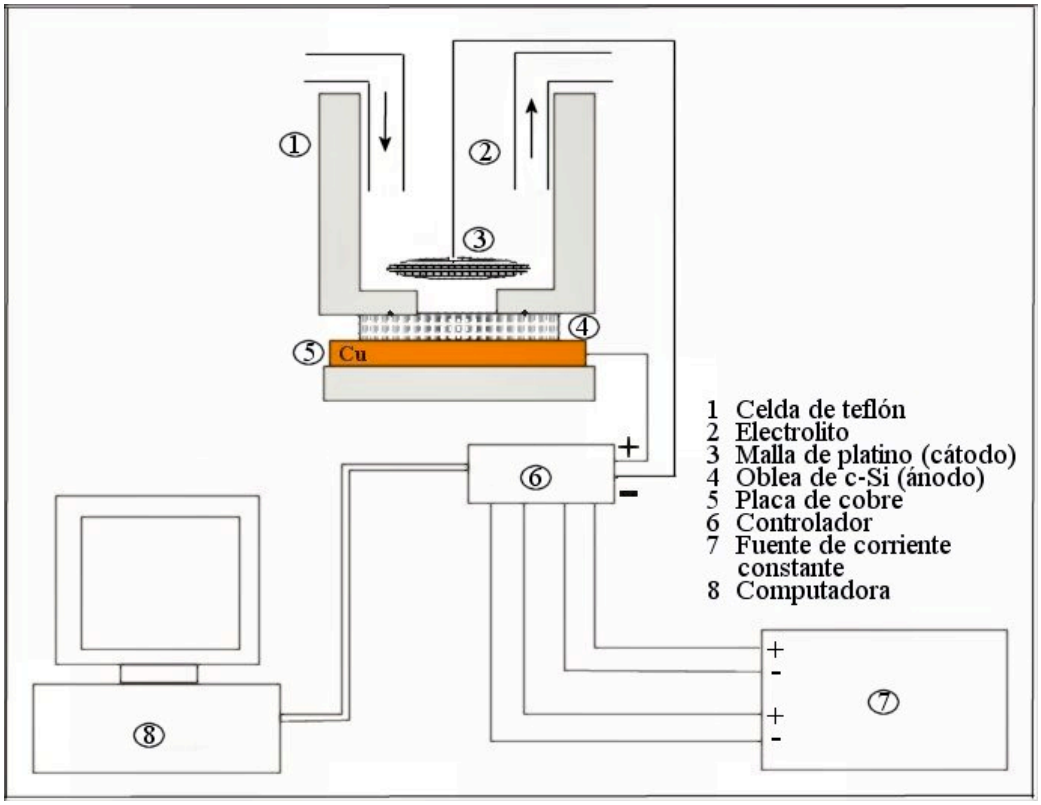


Figura 2: Diagrama del sistema experimental empleado en la producción de silicio poroso

Filtro para luz solar

Un cristal fotónico es un material que permite controlar cómo se propaga la luz dentro de sí mismo. La idea básica en el diseño de estos materiales[8] consiste en crear estructuras que puedan controlar a los fotones en la misma forma en que los semiconductores y los dispositivos construidos con ellos, que permiten manipular a los electrones. En el caso de los semiconductores lo que permite controlar a los electrones es la estructura cristalina de los átomos, que da

origen a un potencial periódico que produce la estructura de bandas, con estados de energía prohibidos y permitidos para los electrones en el semiconductor. En el caso fotónico para manipular la luz se diseña una estructura cristalina en la que el índice de refracción varía de forma periódica.
Cuando ondas electromagnéticas de período espacial similar al periodo de la red entran al cristal, experimentan interferencia y no se propagan en él; ondas electromagnéticas de diferente longitud de onda no interaccionan con la red y se propagan libremente. En conjunto las frecuencias que pueden y no propagarse constituyen la estructura fotónica de bandas del cristal.
En el intervalo de la luz visible la constante de la red del cristal debe ser del orden de 400 a 700 nanómetros para producir interferencia. La tecnología para fabricar objetos de este tamaño existe y es la que se utiliza para fabricar los cristales fotónicos. Pueden construirse unidimensionales, bidimensionales y tridimensionales, según si el índice de refracción varía periódicamente en una, dos o tres dimensiones. Los dos primeros casos poseen una banda fotónica incompleta, lo que significa que el intervalo de frecuencias prohibidas estará solo en la dirección de periodicidad. A diferencia de estos dos un cristal fotónico tridimensional posee una banda fotónica completa lo cual quiere decir que las frecuencias de luz dentro de la banda prohibida se reflejarán y no se propagarán a través de él, en ninguna dirección y sin importar el ángulo de incidencia.
Normalmente el cristal se hace de algún material dieléctrico transparente e independientemente de si es unidimensional, bidimensional o tridimensional, existen cuatro dispositivos básicos que pueden construirse aprovechando su banda fotónica: espejos, filtros, microcavidades y guías de onda.
Vamos a concretarnos en este ejemplo sólo a los espejos y filtros unidimensionales (ver figura 3). Un cristal fotónico se comporta como espejo para las frecuencias dentro de su brecha fotónica, ya que refleja todas estas. Como la brecha puede hacerse más grande, pequeña o recorrerse, un espejo de este tipo puede diseñarse para ser selectivo con las frecuencias que refleja. El cristal se comporta como filtro para las frecuencias de las bandas permitidas ya que las deja pasar.

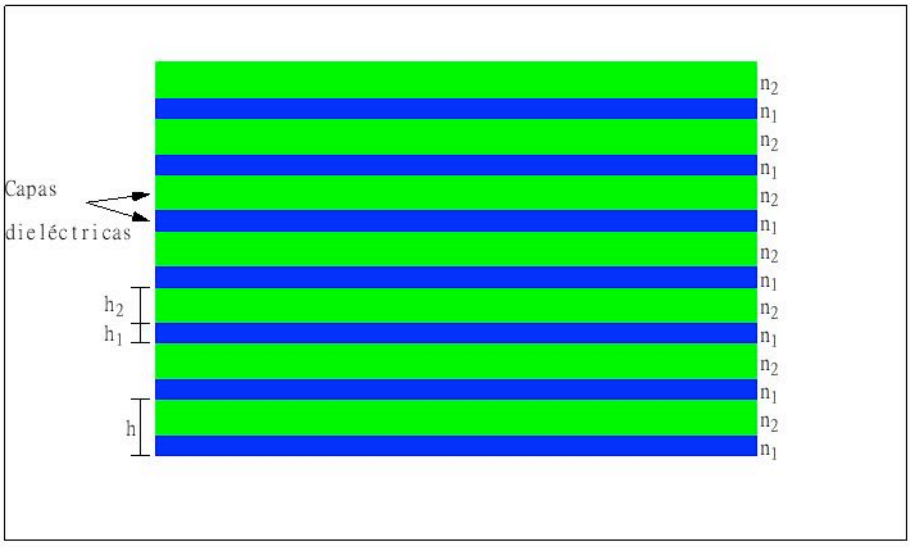


Figura 3. Cristal fotónico unidimensional o multicapa periódica.

En este estudio se buscó el ajuste a un espectro de transmitancia experimental obtenido a 13 m bajo el nivel del mar[10] que se muestra en la figura 4. Se trata de construir un filtro para reproducir el ambiente marino en el que se desarrollan ciertas especies, en particular hay interés en el camarón ornamental, que pierde su coloración en un ambiente luminoso diferente.

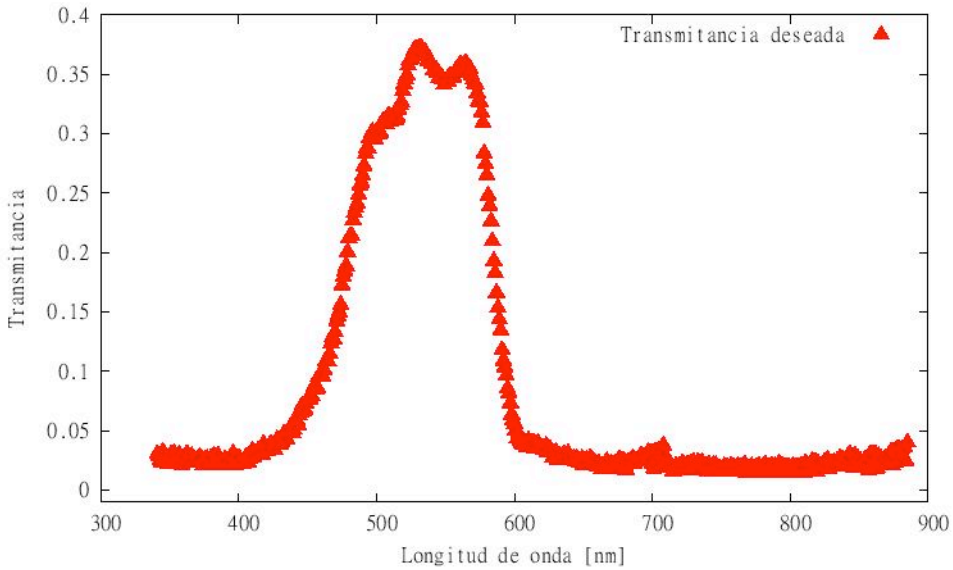


Figura 4. Espectro de transmitancia experimental deseado obtenido a 13 metros de profundidad en el mar.

Para diseñar la multicapa[9] se utilizó el método de la matriz de transferencia[11]. Se consideró una multicapa periódica de un cuarto de onda, con incidencia normal. Se demostró por qué ésta tiene una alta reflectividad para ciertas longitudes de onda y como puede utilizarse para construir un filtro. En el modelo matemático se consideró una multicapa transparente sin absorción y con índices de refracción constantes. En realidad el silicio poroso sí presenta absorción y su índice de refracción varía con la longitud de onda en el intervalo donde está el espectro que se quiere obtener, pero en el rango buscado la aproximación resultó buena para obtener la transmitancia deseada, al ajustar el número de periodos de la multicapa.

Por otra parte, la dependencia del índice de refracción con la longitud de onda hace que las bandas de reflectancia de una multicapa sean más estrechas que las obtenidas con un índice de refracción constante, como en el modelo arriba mencionado. Para incluir este efecto se utilizó el método de ampliar la banda de reflectancia sobreponiendo varias individuales en una sola. Para que cada banda individual fuera lo más ancho posible se escogió un contraste grande entre índices de refracción, se eligió el índice de refracción mayor $\mathbf{n}_1 = 2.4$ y el índice de refracción menor $\mathbf{n}_2 = 1.2$.

Con todas estas consideraciones se elaboró un programa de cómputo basado en la matriz de transferencia que sirvió para calcular la transmitancia de una

multicapa. Con el programa, se diseñó un filtro que reprodujo lo mejor posible, dentro de las limitaciones del modelo la transmitancia en el mar. Con base en este diseño se fabricó una multicapa de silicio poroso y se midió su transmitancia con buenos resultados.
Como era de esperarse, debido a que en la matriz de transferencia no se incluyeron todas las características del silicio poroso, hay diferencias entre la transmitancia obtenida con el programa de cómputo, la del filtro y la marina.
Sin embargo, los espectros marino y del filtro fabricado son lo suficientemente cercanos como para hacerlos coincidir con los ajustes mencionados anteriormente, como se puede ver en la figura 4.

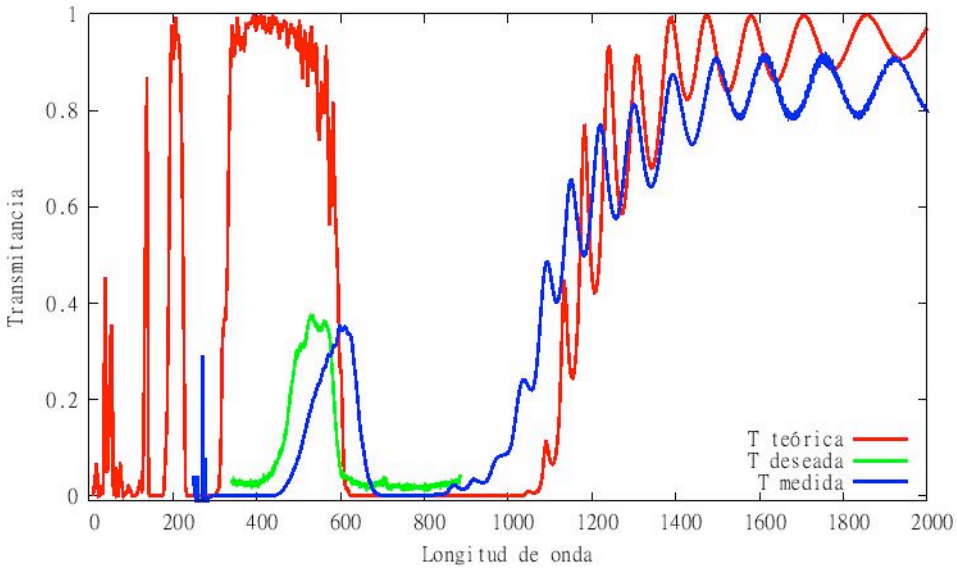


Figura 4. Comparación entre espectros de transmitancia teórico (en rojo), el que se quiere obtener (en verde) y el medido de una multicapa de silicio poroso construida con los parámetros de la matriz de transferencia.

A futuro será muy interesante la investigación experimental que se puede realizar con este filtro, acerca de la influencia que tienen las condiciones luminosas del medio en el crecimiento y la coloración de los camarones.

Conclusiones

Hemos dado un ejemplo de cómo el silicio todavía tiene sorpresas que ofrecer en las nuevas vertientes tecnológicas: la fotónica y en general la nanotecnología. Además estas aplicaciones pueden servir para motivar a que la cultura científica de nuestra sociedad incluya conceptos de mecánica cuántica.

Bibliografía

1.- Capítulo de las Notas de la XVI Escuela de Verano en Física (Cuernavaca, julio-agosto, 2009), El silicio en el siglo XXI: la nanotecnología y los dispositivos fotónicos, J. Tagüeña-Martínez, R. Nava y J.A. del Río, Editores Rocío, Jáuregui, José Recamier y Manuel Torres, impreso por el Instituto de Física y el

Instituto de Ciencias Físicas de la UNAM y la Universidad Autónoma del Estado de Morelos (2009).
2.- En busca del gato de Schrödinger, John Gribbin, Biblioteca Científica Salvat (1986).
3.- Capítulo de las Notas de la Escuela de Verano "La Visión Molecular de La Materia", Los Semiconductores, J. Tagüeña-Martínez, Editor José Recamier, impreso por la Universidad Autónoma del Estado de Morelos (1992).
4.- Asómate a la Materia ¿Qué es un semiconductor? Ciencia de Boleto, Julia Tagüeña, UNAM ISBN 970-32-2071-1 (2005).
5.- Capítulo de las Notas de la II Escuela de Verano "La Visión Molecular de la Materia, El Silicio: de los Cristales a los Poros", J. Tagüeña, M. E. Tejeda y E. J. Lugo, Editor José Recamier, impreso por la Universidad Autónoma del Estado de Morelos (1995).
6.- Capítulo de las Notas de la XVI Escuela de Verano en Física (Cuernavaca, julio-agosto, 2008), La nanotecnología: Silicio Poroso, J. Tagüeña-Martínez, R. Nava y J.A. del Río, Editores José Recamier y Manuel Torres, impreso por el Instituto de Física y el Instituto de Ciencias Físicas de la UNAM y la Universidad Autónoma del Estado de Morelos (2008).
7.- Photonics and Lasers, an introduction, Richard S. Quimby, USA, John Wiley and sons (2006).
8.- John D. Joannopoulos, Steven G. Johnson, Joshua N. Winn, Robert D. Meade Photonic Crystals Molding the Flow of Light Second edition Princeton University Press 2008.
9.- Jazael Gómez Ocampo, "Estructuras de cristales fotónicos de silicio poroso", Tesis de licenciatura Universidad Autónoma del Estado de Morelos, enero 2012.
10.- Fernando Sosa Montemayor, "Estudio de un sistema de concentración solar para sustitución de iluminación convencional en estanques de cultivo de especies marinas", Tesis de doctorado UNAM 2011.
11.- Max Born, Emil Wolf. Principles of Optics 7th edition. Cambridge University Press, 1999.