UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS

Métodos computacionales de búsqueda de comunidades funcionales en redes de regulación genética.

TESIS
QUE PARA OPTAR POR EL GRADO DE:
Doctor en Ciencias

PRESENTA:
M. en C. Sergio Antonio Alcalá Corona

TUTOR PRINCIPAL
Dr. Enrique Hernández Lemus
Instituto de Ecología

COMITÉ TUTOR
Dr. Gustavo Martínez Mekler
Centro de Ciencias Genómicas

Dr. Octavio Miramontes Vidal,
Instituto de Ecologıa

CIUDAD UNIVERSITARIA, ABRIL, 2018

JURADO ASIGNADO:

Presidente:	Dr. Pedro Miramontes Vidal
Secretario:	Dr. Enrique Hernández Lemus
Vocal:	Dr. Pablo Padilla Longoria
Suplente:	Dr. Francisco Fernández de Miguel
Suplente:	Dr. Arturo Carlos Becerra Bracho

La tesis se realizó en: Instituto Nacional de Medicina Genómica (INMEGEN).

TUTOR DE TESIS:

Dr. Enrique Hernández Lemus
Instituto de Ecología

Universidad Nacional Autónoma de México

DOCTORADO EN CIENCIAS BIOMÉDICAS

Métodos computacionales de búsqueda de comunidades funcionales en redes de regulación genética.

T E S I S

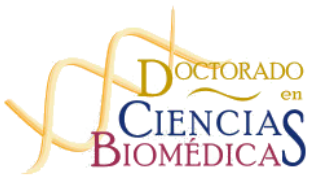
QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:

M. en C. Sergio Antonio Alcalá Corona

TUTOR:
Dr. Enrique Hernández Lemus
Instituto de Ecología



Ciudad Universitaria, Abril, 2018

***A mi familia**:*
***Silvia Ivonne**, mi esposa y el amor de mi vida;*
*a mis padres, **Guadalupe y Antonio**, quienes me la dieron;*
*y **Gustavo**, mi hermano quien siempre me ha acompañado en ella.*

Resumen

EL presente trabajo recoge y resume los principales alcances del proyecto de doctorado intitulado: “Métodos computacionales de búsqueda de comunidades funcionales en redes de regulación genética”.

El proyecto pretende proponer una metodología computacional matemáticamente formal y robusta, para detectar y estudiar *módulos biológicamente funcionales* en *Redes de Regulación Genética* de gran escala, inferidas a partir de datos experimentales. En particular, de datos de genoma completo en humano de biopsias de cáncer de mama, obtenidos de bases de datos publicas.

Comenzaremos con una reseña de los enfoques basados en redes aplicados a la biología molecular, particularmente a las redes de regulación genética. Asimismo, se mostrarán los recientes aportes de la hoy llamada *Ciencia de Redes* a la Biología y a la Medicina.

Posteriormente se hará una exposición breve de las métodos matemáticos para analizar la estructura (topología) de las *Redes Complejas*. Asimismo se detallarán los aspectos formales de la modularidad en redes y se presentará una una breve revisión de los diferentes métodos para identificar módulos derivados de la teoría de Redes Complejas.

Así entonces, se presentará nuestra propuesta metodológica para la identificación de módulos funcionales en redes de regulación genética, construida a partir de los métodos de modularidad en redes. Dicha propuesta consiste en en implementar un algoritmo de detección de módulos basado en el flujo de información sobre la red. Al obtener dichos módulos se realiza un análisis de enriquecimiento funcional estadístico sobre los mismos.

Las principales ventajas de esta metodología se centran en el poder analizar propiedades de gran escala de redes complejas construidas a partir de datos genómicos obtenidos de experimentos. Esto logra una reducción de dimensionalidad en el análisis de las interacciones entre genes, pasando de tener una enorme cantidad de datos que representan la expresión del genoma completo de cierto fenotipo, a una descripción

modular de un modelo matemáticamente robusto de una coregulación concertada. Lo anterior permite obtener indicios acerca de la desregulación del programa transcripcional en algún fenotipo (de tejido enfermo, por ejemplo) con el propósito de postular posibles blancos claves como candidatos experimentales.

Por otro lado, la principal desventaja de este enfoque es que justamente solamente proporciona inferencias teóricas a partir de hipótesis bien determinadas. Sin embargo la realidad biológica puede ser diferente y hay que realizar experimentos para poder corroborar dichas predicciones, las cuales pueden no ser totalmente acertadas.

Por último, se presentarán los resultados de aplicar esta metodología a diferentes casos de estudio, que consisten en diferentes redes construidas a partir de datos provenientes de diferentes bases de datos públicas. En general, el resultado más importante de nuestro estudio, es poder asociar funciones biológicas bien definidas, a pequeños conjuntos de genes detectados en la red a partir de la estructurar topológica de la misma, con una confianza estadística grande. Dicha estructura, está determinada totalmente a partir de modelar matemáticamente la regulación conjunta de todos los genes del genoma, a partir de el patrón de expresión medido experimentalmente en un fenotipo particular.

Los casos de redes abordados son los siguientes. En primer lugar, una red de blancos transcripcionales del *factor de transcripción MEF2C* , en el cual, fue posible encontrar grupos de genes (comunidades o módulos) asociados a procesos biológicos relacionados con señalización celular, expresión génica, ciclo celular y metabolismo y transporte. Los segundos casos de estudio, se refieren a *redes transcrispionales* inferidas a partir de datos genómicos en muestras de *cáncer de mama*. Entre los principales resultados obtenidos en este caso destacan que la estructura modular de las redes inferidas para diferentes subtipos moleculares de cáncer de mama es diferente en cada uno de estos, asimismo los patrones de regulación de diferentes moléculas asociados a los módulos son diferentes aunque se conservan entre los subtipos moleculares de cáncer y están asociados estadísticamente a diferentes funciones biológicas. Así también, estás redes presentan una sub-estructurara modular jerárquica interesante de estudiar pues puede develar funciones biológicas mucho más particulares de la coregulación genética en cáncer de mama, como se demuestra en la red asociada al subtipo *Her2+*.

Finalmente, a partir de describir nuestra metodología y mostrar los resultados obtenidos se expondrán las conclusiones del proyecto doctoral.

Índice general

Índice de figuras IX

Índice de tablas XXI

1. Ciencia de Redes aplicada a la Biología. 1
- 1.1. **El programa de regulación transcripcional genético.** 2
  - 1.1.1. El proceso de Regulación Transcripcional Genética. 2
  - 1.1.2. Redes de Regulación genética clásicas. 5
- 1.2. **Bioinformática y Biología de Sistemas.** 12
  - 1.2.1. Herramientas para el análisis de datos genómicos. 13
  - 1.2.2. **Biología de Sistemas.** 14
  - 1.2.3. Inferencia de Redes de coexpresión de genes. 18
- 1.3. **Redes *Complejas* aplicadas a la biología y enfermedades.** 19
  - 1.3.1. Biología de Red (*Network Biology*). 20
  - 1.3.2. Redes aplicadas al estudio de enfermedades. 23
    - 1.3.2.1. Aplicaciones al cancer. 32
- 1.4. **Objetivo y planteamiento del proyecto de investigación.** 34
  - 1.4.1. Pregunta de Investigación y objetivos. 35
  - 1.4.2. Panorama de la Tesis doctoral. 35

2. Redes Complejas. 37
- 2.1. **Definiciones generales con base en la teoría de gráficas.** 38
  - 2.1.1. Gráficas dirigidas y pesadas. 39
    - 2.1.1.1. Subgráficas y otros tipos de gráficas relevantes. 40
    - 2.1.1.2. *Redes Complejas.* 41
  - 2.1.2. Representación y algunas propiedades matriciales de redes. 41
    - 2.1.2.1. Matriz de Adyacencia. 41
    - 2.1.2.2. Propiedades espectrales. 43
    - 2.1.2.3. Matriz Laplaciana. 43
- 2.2. **Centralidades básicas y caracterización de redes complejas.** 44
  - 2.2.1. Conectividad y distribución de vecinos. 45
  - 2.2.2. Distancia mínima promedio entre nodos y *efecto de mundo pequeño.* 46
    - 2.2.2.1. Efecto de mundo pequeño. 48

2.2.3. Coeficiente de agrupamiento (*Clustering Coefficient*). 49
2.2.4. Distribución de Grado **P(k)**. 51
2.2.5. Otras centralidades y propiedades de Redes Complejas. 53
2.2.5.1. Centralidades de nodo. 53
2.2.5.2. Propiedades de Redes. 57
2.3. **Modelos Redes Complejas: Topologías y distribuciones de de grado**. 60
2.3.1. Redes Aleatorias y Topología de Poisson (Modelo de Erdös-Rényi). 61
2.3.1.1. Percolación y la *Isla Gigante* 62
2.3.1.2. Distancia entre nodos y Clustering Coefficient 65
2.3.2. Redes de mundo pequeño (Modelo de Watts y Strogatz). 66
2.3.3. Crecimiento en redes mediante enlace preferencial y topolgía de Libre de Escala (Modelo de Albert-Barabasi). 69
2.3.3.1. Modelo de Barabasi-Albert. 70
2.4. **Estado del Arte en Redes Complejas**. 73
2.4.1. Redes Multicapa (Multilayer Networks). 75

3. Modularidad y métodos computacionales de detección de comunidades en redes complejas. 77
3.1. **Modularidad y estructura modular en Redes Complejas.** 78
3.1.1. Ejemplos de Redes reales con estructura modular. 78
3.1.1.1. Modularidad en Redes Sociales. 79
3.1.1.2. Modularidad en Redes de colaboración y citas de artículos. 81
3.1.1.3. Modularidad en Redes Informáticas. 83
3.1.2. Modularidad en Redes biológicas. 84
3.1.2.1. Redes de Interacción de Proteínas. 85
3.1.2.2. Redes Metabólicas y genéticas. 88
3.1.3. Relevancia y aplicaciones de la detección de comunidades en redes complejas. 89
3.2. **Elementos de detección de Comunidades** 91
3.2.1. Definiciones formales de Modulo. 92
3.2.2. Complejidad Computacional del problema. 93
3.2.3. Medida de *Modularidad* (evaluación de particiones). 93
3.2.4. Consideraciones especiales de la Estructura Modular. 96
3.2.4.1. Redes Dirigidas y Pesadas. 96
3.2.4.2. Redes Multipartitas. 97
3.2.4.3. Jerárquica y superposición de módulos. 97
3.3. **Métodos de detección de comunidades.** 99
3.3.1. Técnicas Tradicionales: Agrupación jerárquica y k-means clustering. 100
3.3.2. Algoritmos Divisivos 102
3.3.2.1. El algoritmo de Girvan-Newman. 103
3.3.2.2. Otros métodos divisivos. 104
3.3.3. Algoritmos Espectrales 105

3.3.3.1. Bisección Espectral (spectral bisection method.) 106
3.3.3.2. El algoritmo de Donetti y Muñoz. 106
3.3.3.3. Redención de los métodos espectrales. 108
3.3.4. Algoritmos basados optimización de Modularidad. 109
3.3.4.1. El algoritmo de Clauset-Newman. 109
3.3.4.2. El método de Louvain (Blondel *et al.*) 110
3.3.4.3. Templado simulado (Simulated annealing). 110
3.3.4.4. Optimización espectral (Spectral optimization). 111
3.3.4.5. Límites de la optimización de Modularidad. 112
3.3.5. Algoritmos Dinámicos. 113
3.3.5.1. Modelos de Spin. 113
3.3.5.2. Label Propagation Method 114
3.3.5.3. Sincronización 115
3.3.5.4. Algoritmos basados en Caminatas Aleatorias 115
3.3.6. Otras metodologías 120
3.3.6.1. OSLOM 120
3.3.6.2. Stochastic Block Models (SBM) 120
3.3.6.3. Percolación de Clique (Clique Percolation). 122
3.4. **Métodos de prueba de algoritmos de detección de comunidades.** 122
3.4.1. Redes Reales 123
3.4.1.1. Red del Club de Karate de Zachary. 124
3.4.1.2. Otras Redes Reales. 125
3.4.2. Redes de prueba (benchmarks) generadas computacionalmente. . 127
3.4.2.1. Modelo de *partición-l plantada*. 127
3.4.2.2. Prueba Girvan-Newman. 128
3.4.2.3. Prueba LFR. 129
3.4.3. Medidas de comparación de particiones. 131
3.4.3.1. Fracción de nodos correctamente clasificados. 131
3.4.3.2. Información Mutua Normalizada. 132
3.4.4. Comparación de algoritmos 134

4. Detección de módulos biológicamente funcionales. **141**
4.1. **Propuesta Metodológica.** 142
4.1.1. Detección de Comunidades e Infomap. 144
4.1.1.1. Codificación de una caminata aleatoria infinita en la red. 144
4.1.1.2. Teoría de la información y la *MapEquation* 146
4.1.1.3. Minimización de $L(M)$ y refinamiento por *Templado Simulado* 149
4.1.1.4. Detección de Submodularidad Jerárquica. 150
4.1.2. Análisis de Enriquecimiento estadístico. 152
4.1.2.1. Análisis de Sobre-Representación y Prueba Hipergeométrica. 153
4.1.3. Análisis de Expresión Diferencial. 155

4.2. **Casos de Estudio.** . . . 155
4.2.1. Red transcripcional de MEF2C. . . . 155
4.2.1.1. Construcción de la red a partir de Análisis de sitio de unión de *Factor de Transcripción* (TFBS). . . . 156
4.2.1.2. Análisis de Enriquecimiento funcional. . . . 157
4.2.1.3. Construcción de *Modelo Nulo.* . . . 158
4.2.2. Redes transcripcionales de subtipos de cáncer de mama. . . . 158
4.2.2.1. Cáncer de mama: una enfermedad heterogénea. . . . 158
4.2.2.2. Inferencia de redes a partir de microarreglos de cáncer de mama. . . . 160
4.2.2.3. Análisis de Enriquecimiento funcional. . . . 162
4.2.3. Submodularidad jerárquica en la red transcripcional del subtipo Her2 de cáncer de mama. . . . 164

5. Resultados y Conclusiones. **165**
5.1. **Resultados de los casos de estudio.** . . . 165
5.1.1. Red transcripcional de MEF2C. . . . 165
5.1.1.1. La estructura modular revela *sub-redes* que participan en procesos biológicos específicos. . . . 166
5.1.1.2. La estructura modular es validada por un *modelo nulo.* 173
5.1.2. Redes transcripcionales de subtipos de cáncer de mama. . . . 174
5.1.2.1. La estructura modular es específica para cada subtipo y está asociada a procesos biológicos específicos. . . . 174
5.1.2.2. El módulo asociado a COL5A2 está presente en cada subtipo molecular. . . . 176
5.1.2.3. Análisis de los módulos del subtipo Basal y la disminución de la *apoptosis* mediante el módulo de *PSMB9.* . . . 179
5.1.3. Submodularidad jerárquica en la red transcripcional del subtipo Her2 de cáncer de mama. . . . 185
5.1.3.1. Enriquecimiento y análisis de expresión diferencial, revela la función biológica en submódulos. . . . 187
5.1.3.2. Submódulos CNR2 y la membrana plasmática. . . . 190
5.1.3.3. Submódulos de LCK, respuesta viral y respuesta inmune celular. . . . 191
5.2. **Discusión y Conclusiones.** . . . 194
5.2.1. Comunidades Funcionales en la red de transcripción de MEF2C. 195
5.2.2. Comunidades Funcionales en redes de cáncer de mama. . . . 196
5.2.3. Conclusiones Generales. . . . 198

Bibliografía **201**

Índice de figuras

1.1. Esquema de regulación transcripcional *positiva* y *negativa*. 3
1.2. Esquema de funcionamiento de factores de transcripción. 4
1.3. Modelación de la regulación de algunos genes mediante una gráfica pequeña. Las flechas rojas representan una *regulación positiva* (promoción) de la transcripción del gen de donde parten hacía donde llega la flecha. Las flechas azules representan una *regulación negativa* (o inhibición) de la transcripción. Cuando una flecha entra y sale del mismo gen (nodo) representa una autoregulación. 5
1.4. Red de interaccción de proteinas en Sacaromise cerevisiae. 6
1.5. Comparación de algunos genes involucrados redes de regulación en el desarrollo corporal en diferentes organismos. 7
1.6. Comparación de módulos en redes de regulación genética asociadas al desarrollo en mosca (*Drosophila melanogaster*) y ratón (*Mus musculus)* . 8
1.7. Tabla de verdad: función de actualización de una red booleana. 9
1.8. Ejemplos de aplicación del modelo de Kauffman. 10
1.9. Ejemplos de aplicación del modelo de Kauffman. 11
1.10. Modulos en la red de regulación genética subyacentes a la determinación de la célula-destino del órgano floral primordial asociada al desarrollo de flores en *Arabidopsis thaliana*. Figura tomda de Alvarez-Buylla *et al.*, [42] 12
1.11. Microarreglo de expersión. 14
1.12. Estudio en diferentes "escalas *omicas*" de interacciones entre moléculas. . 15
1.13. La *Biología de Sistemas* estudia a los sistemas biológicos en diferentes escalas, a partir de las grandes cantidades de datos *omicos* disponibles hoy en día. 16
1.14. Esquema representativo que muestra el caracter integrativo, interdiciplinario y de retroalimentación de la *Biología de Sistemas*. 17
1.15. Inferencia de Redes transcripcionales de *co-expersión* a partir de *microarreglos*. 18
1.16. Esquema de como las redes celulares pueden ayudar a entender las relaciones genotipo-fenotipo. Las perturbaciones en los sistemas biológicos pueden modelarse como cambios estructurales en las redes complejas celulares que forma los genes y sus productos. 21

1.17. Una red compleja. A) Topología libre de escala, B) estructura modular y C) jerárquica 22
1.18. Redes complejas en diferentes escalas aplicadas a la biología y la salud. . 23
1.19. *Diseasome*: las enfermedades humanas, forman una red bipartita (sección 2.1.1.1, capitulo 2) en la que dos enfermedades están conectadas si comparten al menos un gen. Las proyecciones se muestran a la derecha (red de enfermedades) e izquierda (red de genes). 24
1.20. *Interactoma*: Diversas redes biologicas de diferentes tipos, en diversos organismos como levadura, gusano, mosca y planta, interactuan de manera conjunta. 26
1.21. *Módulo de enfermedad*. (a) Los genes asociados a enfermedades (nódulos azules), tienden a co-localizarse en la red de interacción proteína-proteína humana (nódulos blancos), formando un módulo (óvalo azul). (b) Un módulo de enfermedad real de pacientes alérgicos. 28
1.22. *DIseAse MOdule Detection (DIAMOnD)*. Proteínas asociadas con el mismo fenotipo tienden a localizarse en vecindarios específicos del Interactoma. 29
1.23. Módulos Funcionales: Procesos biológicos enriquecidos estadísticamente en módulos topológicos de una red de interacción de proteínas, pueden mapear a módulos en una red de enfermedades. 30
1.24. Integración bajo el enfoque de *Biología de Sistemas* de diferentes capas de redes biológicas (y sociales) en varias escalas. Un claro ejemplo de la aplicación de la *Ciencia de Datos* y la *Ciencia de Redes* a los sistemas biológicos 31
1.25. *Hallmarks of Cancer*: Los *sellos distintivos del cáncer* muestran lo compleja y heterogénea que es ésta enfermedad. 32
1.26. Redes asociadas inferidas a partir de biopsias de cáncer de mama. Se muestran las diferencias topológicas de las redes asociadas a los cuatro subtipos moleculares y la red de tejido sano. 33

2.1. Un ejemplo sencillo de una gráfica (*binaria*) con siete vértices y siete aristas. 39
2.2. **Gráficas dirigidas y pesadas**. **A)** Una gráfica simple dirigida. **B)** Una gráfica pesada. **C)** Un gráfica dirigida y pesada. 40
2.3. **Arboles**. **A)** Ejemplo de un árbol simple. **B)** Un árbol simple dirigido. 41
2.4. **Ejemplos de gráficas bipartitas**. **A)** Nótese la separación en dos subconjuntos *disjuntos* (clases). **B)** Las dos proyecciones únicas de una red bipartita. La parte central de esta figura muestra una red bipartita con cuatro vértices de un tipo (círculos abiertos etiquetados de A a D) y siete de otro (círculos rellenos, de 1 a 7). En la parte superior e inferior, se muestran las proyecciones de los dos conjuntos de vértices. 42
2.5. **Matriz laplaciana de una red**. **A)** Una pequeña red no dirigida **B)** Matriz de Adyacencia $\mathbb{A}$ que describe la conectividad de red de la red en el panel A; **C)** Definición de la matriz laplaciana de una red; **D)** matriz laplaciana $\mathbb{L}$ de la red en el panel A. 44

2.6. **Camino más corto en un componente conexo de una red**. Aunque existen varios caminos para llegar de i a j, el camino más corto está indicado con líneas gruesas. Figura tomada de las notas de Maximino Aldana.. 47
2.7. **Clustering Coefficient**. **A)** El *Clustering Coefficient* del nodo x es la relación entre los triángulos que si están formados con el nodo x y todos los posibles triángulos que se pueden formar con el nodo x; **B)** Diferentes valores de *Clustering Coefficient* para el nodo i. 50
2.8. **Diferencias en la topología de redes**. A la izquierda se muestra una *red aleatoria* con topología exponencial (o *Topología de Poisson* y su distribución de grado (abajo). A la derecha se muestra una red con *Topología Libre de Escala* y (abajo) su distribución de grado graficada en $log(P(x))$ vs. $log(k)$. 52
2.9. **Centralidad de intermediación (*betweenness centrality*)**. Se muestra un nodo con alta intermediación en una red personajes de la Florencia renacentista. 54
2.10. **Esquemas representativos de la centralidad de PageRank**. El tamaño de los nodos es proporcional a su *PageRank* y este a su vez toma en cuenta las conexiones entrantes de ponderando las de mayor *PageRank*. Nótese como el nodo C del panel A, tiene un PageRank alto a pesar de solo tener conexión con el nodo B. 56
2.11. **Varios *componentes* de una red**. Si no existe un camino entre algún par de nodos, la red se divide en *componentes*, a veces llamados islas. Figura tomada de las notas de Maximino Aldana.. 58
2.12. **Distribuciones de grado para el Modelo de Erdös-Rényi**. Se muestran tres diferentes distribuciones de grado para redes generadas con el *Modelo de Erdös-Rényi*: $z = \langle \mathbf{k} \rangle = 5$, $\langle \mathbf{k} \rangle = 10$, $z = 20$. 62
2.13. **Redes generadas con el *Modelo de Erdös-Rényi***. Se muestran tres diferentes instancias de redes generadas con el *Modelo de Erdös-Rényi* para varios valores de $z = \langle \mathbf{k} \rangle$. 63
2.14. **Transición de Fase el *Modelo de Erdös-Rényi***. Se muestran la probabilidad de que todos los nodos de la red pertenezcan al componente conexo más grande (gigante) en función de $z = \langle \mathbf{k} \rangle$. Se muestra que a partir de $z = \langle \mathbf{k} \rangle = 1$ existe una *transición de fase* y el sistema percola hacia un sólo *componente gigante*. Figura tomada de las notas de Maximino Aldana.. 64
2.15. ***Modelo de Watts-Strogatz***. Se muestra el procedimiento bajo el cual se construyen redes usando el *Modelo de Watts-Strogatz* al reconectar los nodos con una probabilidad p_r. Al aumentar la probabilidad de reconexión p_r se nota una transición de la red en un régimen regular (*lattice*) hacia una red aleatoria tipo *Erdös-Rényi*. 67

2.17. **Clustering Coefficient y Shortest Path Length en el *Modelo de Watts-Strogatz***. Se muestran la gráfica de $\langle l \rangle$ y $\langle C \rangle$ vs. la probabilidad de reconexión p_r. La zona gris marcada con una flecha muestra un régimen en el que el *efecto de mundo pequeño* y un coeficiente de agrupamiento alto coexisten en el mismo sistema, lo cual es una característica de las redes reales. 67
2.16. **Cambio en la distribución de grado en el *Modelo de Watts-Strogatz***. La distribución de grado pasa de una *delta de Dirac* (todos los nodos con exactamente el mismo grado) a una distribución de una red aleatoria *Erdös-Rényi* (la mayoría de los nodos tienen un grado muy similar, cercano al promedio. 68
2.18. **Comparación de regímenes de redes**. Se muestra los regímenes de redes entre los que se encuentran las redes reales: redes aleatorias (*Modelo de Erdös-Renyí*) y redes ordenadas (*lattices*). A pesar de no ser perfectos tanto los modelos de *Barabasi-Albert* como de *Watts-Strogatz* se acercan a las redes reales. 68
2.19. **Red Libre de Escala generada con el *Modelo de Barabasi-Albert***. Nótese la presencia de nodos *hubs* que acaparan muchas conexiones, mientras la mayoría de los nodos sólo tiene pocas conexiones. 70
2.20. **Comparación de distribuciones de grado respecto del *Modelo de Barabasi-Albert***. Se muestran comparaciones con las distribuciones de grado del *Modelo de Erdös-Renyí* (Poisson) y Topología y exponencial. . 71
2.21. **Comparación de Topologías de Red**. Se muestra la diferencia entre 3 principales topologías de red en términos de su distribución de grado y de Clusetering Coefficient. 72
2.22. **Comparación de Modelos de Red**. Arriba: una red aleatoria generada con el *Modelo ER*, en medio: una red aleatoria generada con el *Modelo WS* y abajo: una red aleatoria generada con el *Modelo BA*. 73
2.23. **Ejemplos de Redes Multicapa**. Derecha: una *red Multilayer* (los nodos tienen conexiones entre ellos en una capa y entre los de las otras capas). Izquierda: una *red Multiplex* (los nodos son las mismas entidades por lo que están conectados con ellos mismos en las diferentes capas, pero tienen diferentes interacciones con los demás nodos en cada capa). . . . 75
2.24. Las Redes Multicapa parecen ser un marco teórico prometedor para modelar redes de sistemas biologicos moleculares a diferentes escalas. 76

3.1. **Ejemplo esquemático simple de una pequeña gráfica con estructura modular**. Se muestran tres conjuntos de nodos (rojo, verde y azul) con una mayor densidad de aristas entre ellos respecto a los otros conjuntos de nodos. 78
3.2. **Red de amistades de niños en una escuela de Estados Unidos tomada de un estudio de Moody**. Se puede notar que una de las divisiones de la red es la etnia y otra la división entre la escuela media y la secundaria. 79

3.3. **Ejemplos de modularidad en diferentes redes sociales**. **a**: la famosa red del *Club de Katate de Zachary* 3.4.1.1 con una partición en 4 grupos (diferente a la tradicional; **b**: Red de colaboraciones en el *Santa Fe Institute* **c**: Red de delfines nariz de botella de Lusseau. 80
3.4. **Modularidad en redes de citas**. Mapa de la ciencia de Rosvall y Bergstrom derivado de un análisis de una red de citas que comprende más de 6000 revistas. Los autores usaron su método *Infomap*, el cual cómo veremos más adelante (secciones 3.3.5.4,3.4.4 y 4.1.1) es una de las mejores propuestas para encontrar la estructura modular de una red. Los nodos representan *áreas de conocimiento* (módulos) y las aristas y su grosor el flujo de información (de citación) entre las áreas. 82
3.5. **Estructura modular en una red de interacción proteína-proteína**. Se muestran las interacciones entre proteínas en células cancerosas de rata. Las comunidades, etiquetadas por colores, fueron detectadas con el *Método de Percolación de Clique* de Palla *et al.* (sección 3.3.6.3). 86
3.6. **Ejemplo esquemático de una red con estructura modular jerárquica**. La red consta de cuatro módulos principales, cada uno con 4 submódulos. Los 16 submódulos constan de 32 nodos con grado $k = 64$. 98
3.7. **Comunidades superpuestas (o traslapadas).** En la partición resaltada por los contornos discontinuos, algunos nodos se comparten entre más de una comunidad. 99
3.8. **Particionamiento de red.** El corte muestra la partición en dos grupos de igual tamaño. 100
3.9. **Dendrograma (o árbol jerárquico).** Este dendograma corresponde a la partición de la red del *club de Karate de Zachary* (sección 3.4.1.1) mediante el algoritmo de Girvan-Newman (sección 3.3.2.1). Los cortes horizontales corresponden a las particiones del gráfico en las comunidades. 101
3.10. **Intermediación de arista (edge betweenness).** La intermediación es más alta para aristas que conectan comunidades. La arista gruesa en el centro tiene una intermediación mucho más alta que todas las demás aristas, porque todos los caminos más cortos que conectan los nodos de las dos comunidades lo atraviesan. 103
3.11. **Algoritmo espectral de Donetti y Muñoz.** El nodo i está representado por la i-ésima entrada de los vectores propios de la matriz Laplaciana. En este ejemplo, la red tiene una partición en cuatro módulos, indicados por diferentes símbolos (y colores). Las comunidades están mejor separadas en dos dimensiones (derecha) que en una (izquierda). 107

3.12. **Método de Louvain (Blondel *et al.*).** El diagrama muestra dos iteraciones del método de optimización jerárquica de la modularidad, comenzando por la red de la izquierda. Cada iteración consiste en un paso, en el que cada nodo se asigna a la comunidad (local) que produce el mayor aumento de modularidad, seguido de una transformación sucesiva de los módulos en nodos de una red (pesada) más pequeña, representando el siguiente nivel jerárquico superior. 111
3.13. **Infomap de Rosvall y Bergstrom**. La caminata aleatoria en (**A**) se puede describir como una secuencia de nodos, cada uno etiquetado con *palabras de código* únicas (**B**), o dividiendo la red en regiones y usando *palabras de código únicas* solo para los nodos de la misma región (**C**). De esta forma, la misma *palabra de código* puede usarse para múltiples nodos, a costa de indicar cuándo el caminante aleatorio abandona una región para ingresar a una nueva, ya que en ese caso hay que especificar la palabra de código de la nueva región para ubicar al caminante. La red tiene cuatro comunidades (indicadas por los colores en (C)), y en este caso la descripción tipo mapa de (C) es más compacta que la de (B). Esto se muestra al mirar el código real que se necesita en cualquier caso (parte inferior de las figuras), que es claramente más corto para (C). En (**D**) las transiciones entre los módulos están resaltadas en la *codificación*. 119
3.14. **Stochastic Block Model**. Se muestran las matrices de adyacencia esquemáticas de redes generadas por el modelo para elecciones especiales de probabilidades de arista. Los bloques más oscuros indican mayores probabilidades de arista y, en consecuencia, una mayor densidad de aristas dentro del bloque. (a) Ilustra la estructura modular (o assortative): las probabilidades (densidades de enlace) son mucho más altas dentro de los bloques diagonales que en cualquier otro lugar. (b) Muestra la situación opuesta (estructura disassortativa). (c) Ilustra una estructura núcleo-periferia. (d) Muestra un gráfico aleatorio tipo Erdős y Rényi: todas las probabilidades de arista son idénticas, dentro y entre los bloques, por lo que no hay grupos reales. 121
3.15. **Red del *Club de Karate* de Zachary**. Se muestra la partición estándar en dos comunidades. Los círculos rojos representan una comunidad y los azules la otra. 124
3.16. **Red de equipos de fútbol americano colegial de universidades estadounidenses.** 125
3.17. **Red de delfines de naris de botella estuadiada por Lusseau**. En cuadros de color verde se representa una comunidad y los circulos amarillos la otra. 126
3.18. **Benchmark de Girvan y Newman**. Tres redes generadas con la *prueba GN* que corresponden a: (**a**) $\langle k \rangle_{in} = 15$; (**b**) $\langle k \rangle_{in} = 11$ y (**c**) $\langle k \rangle_{in} = 8$. Nótese que en el caso (c) los cuatro módulos son difíciles de notar. . . . 128

3.19. **Ejemplo de una red con 500 nodos generada con la *prueba LFR*.** Las distribuciones de grado y del tamaño de comunidad siguen leyes de potencia. Este tipo de pruebas se aproximan más fielmente a las redes de mundo real con estructura modular. 130
3.20. **Ejemplo de una red con 500 nodos generada con la *prueba LFR*.** Las distribuciones de grado y del tamaño de comunidad siguen leyes de potencia. Este tipo de pruebas se aproximan más fielmente a las redes de mundo real con estructura modular. 133
3.21. **Tabla comparativa de algoritmos en el estudio realizado por Danon *et al.*.** Se muestra la complejidad algorítmica de cada método. . 135
3.22. **Rendimiento de algoritmos en el estudio realizado por Danon *et al.*.** Para cada método se tiene un grupo de tres columnas ($\langle k \rangle_{out} = z_{out}$ =6, 7 y 8), cuya altura representa la fracción de nodos correctamente clasificados de la *partición encontrada* respecto a la *partición plantada*. . 135
3.23. **Tabla comparativa de algoritmos en el estudio realizado por Lancichinetti *et al.*, prueba LFR.** 136
3.24. **Comparación de métodos bajo la *prueba LFR*.** Se muestra el rendimiento del *algoritmo de Girvan-Newman* (sección 3.3.2.1) para el cual solo se adoptó el tamaño más pequeño, debido al alto tiempo de computo del método. Así también se muestra el rendimiento para el *algoritmo de Donetti y Muñoz* (sección 3.3.3.2) y para el *método de Ronhovde y Nussinov* (sección 3.3.5.1). 137
3.25. Se muestra el rendimiento del *Método de Percolación de Clique (Clique Percolation)* o *Cfinder* de Palla *et al.* (sección 3.3.6.3); el del *algoritmo de Radicchi et al.* (sección 3.3.2.2). Así como para el *algoritmo de Clauset-Newman* (sección 3.3.4.1) y el *Templado simulado (Simulated annealing)* de Guimerá *et al.* (sección 3.3.4.3). 137
3.26. **Comparación de métodos bajo la *prueba LFR*.** Se muestra el rendimiento del *método de Louvain* de Blondel *et al.* (sección 3.3.4.2); el del *Markov Cluster Algorithm* o MCL de Van Dongen (sección 3.3.5.4). Así como para **Infomap** de Rosvall y Bergstrom (sección 3.3.5.4). 138
3.27. **Comparación de Infomap y OSLOM bajo la *prueba LFR*.** El rendimeinto del *Método de Optimización Local de Estadísticas de Orden* (**OSLOM**) de Lancichinetti *et. al.* (sección 3.3.6.1) es comparable con el de **Infomap** de Rosvall y Bergstrom (sección 3.3.5.4). A la izquierda se muestra el rendimiento para diferentes tamaños de sistema correspondientes a la prueba LFR estándar (1000 y 5000 nodos) y rangos de tamaño de comunidad: (**S**) de 10 a 50 nodos y (**B**) de 20 a 100 nodos. A la derecha se muestra el rendimiento para redes muy grandes correspondientes a 50,000 y 100,000 nodos. 139

4.1. Propuesta Metodológica para la búsqueda de módulos funcionales en redes de regulación genética. 143

4.2. **Infomap y la *MapEquation***. **A**: Una caminata aleatoria alcanza una distribución estacionaria de visitas a los nodos en la Red. **B**: La dinámica de esta caminata puede codificarse la *codificación de Huffman*. **C**: Usando una descripción de la caminata más eficiente se pueden localizar estructuras (módulos y/o nodos) con mayor información en la red. **D**: la *mapequation* describe el movimiento entre módulos y dentro de los mismos, mediante la descripción optima de la caminata, lo que descubre los módulos de la red; nótese las transiciones entre los módulos están resaltadas en la *codificación* (parte inferior). 148
4.3. Ejemplo del uso de la *MapEquation* Jerárquica. 151
4.4. **Base de Datos usadas para análisis de enriquecimiento**. Las bases de datos como *Kyoto Encyclopedia of Genes and Genomes* (**KEEG**), *PathwayCommons* (**PC**) o el *Gene Ontology Consortium* (**GO**), tienen anotadas funciones biológicas bien estudiadas para genes. 152
4.5. Factor de Transcripción MEF2C. 156
4.6. Red a de blancos transcripcionales de MEF2C a tres niveles de profundidad. 157
4.7. Esquema que representa los cuatro tipos de cáncer de mama su porcentaje en casos de 159
4.8. Redes asociadas inferidas a partir de biopsias de cáncer de mama. Se muestran las diferencias topológicas de las redes asociadas a los cuatro subtipos moleculares y la red de tejido sano. 161
4.9. **Propuesta Metodológica para la búsqueda de módulos funcionales en redes de regulación genética asociados a los 4 subtipos moleculares de cáncer de mama**. El flujo de trabajo muestra la clasificación en subtipos moleculares de las muestras de biopsias de cáncer de mama; posteriormente la inferencia de redes y la partición en módulos de las mismas; para finalmente realizar un análisis de enriquecimiento el la base de datos Gene Ontology con el objetivo de asociar funciones biológicas a los módulos. 163

5.1. Red transcripcional de interacciones TFBS para el Factor de Transcripción MEF2C y sus blancos hasta el tercer nivel. En esta visualización, el color y el tamaño de los genes se representa según el grado de nodo (número de vecinos conectados a este gen particular): pequeños nodos verdes corresponden a genes apenas conectados; mientras que los nodos rojos y naranjas más grandes representan genes altamente conectados. . 166

5.2. Estructura modular de la red transcripcional de MEF2C (figura 5.1). Los módulos se etiquetan con el nombre del nodo con el mayor PageRank (sección 2.2.5.1) dentro de la comunidad. Los nodos (que representan módulos) se representan de acuerdo con el tamaño de la comunidad y el flujo de información dentro de esa comunidad. En este sentido, los colores más oscuros corresponden a contenidos de información más grandes, mientras que los círculos más grandes representan comunidades más grandes. El grado relativo de flujo de información se representa en el ancho y el color de los enlaces entre módulos. El grosor de los bordes del módulo refleja la probabilidad de que un caminante aleatorio dentro del módulo siga una regulación (borde) a un gen fuera del módulo. Los enlaces ponderados entre comunidades representan *flujo de regulación*, con el color y el ancho de los bordes que indican el volumen de flujo. Por ejemplo, las líneas entre las comunidades JUND y ELF2 indican un flujo de información de regulación entre ellas. Estos enlaces revelan la relación entre las comunidades. 168
5.3. Mapa de calor que representa el enriquecimiento en las vías relacionadas con procesos de *señalización celular*. La intensidad del color es proporcional al $-log(p_v)$. Los puntos más oscuros corresponden a éxitos estadísticamente significativos. 169
5.4. Mapa de calor que representa el enriquecimiento en las vías relacionadas con *ciclo celular*. Los módulos se etiquetan con el nombre de su molécula de con mayor PageRank (sección 2.2.5.1). El recuadro superior izquierdo muestra el histograma de *Z-score* y el color para el *p-value* de los procesos enriquecidos. 170
5.5. Enriquecimiento en las vías relacionadas con *expresión génica*. El recuadro superior izquierdo también describe un dendrograma que muestra distribuciones similares de *p-values* entre los procesos enriquecidos en las comunidades. 171
5.6. Enriquecimiento en las vías relacionadas con *metabolismo y procesos de transporte celular*. 172
5.7. **Panel A.** Modelo nulo construido para la red de MEF2C. **Panel B.** Estructura modular del modelo nulo. 173
5.8. **Módulos en redes para cada subtipo molecular de cáncer de mama**. A) Luminal A; B) Luminal B; C) Her2 + y D) Subtipo basal. Los nodos que pertenecen a una comunidad tienen el mismo color. . . . 175
5.9. **Comunidades COL5A2: los procesos enriquecidos se comparten entre subtipos a pesar de que las composiciones genéticas son diferentes**. **A)** Diagrama de Venn que muestra el número de genes de $COL5A2_{comm}$ para cada subtipo molecular. Tenga en cuenta que solo se comparten 3 genes (*COL5A2*, *THBS2* y *LUM*). **B)** Procesos enriquecidos de $COL5A2_{comm}$ para cada subtipo molecular. 178

5.10. **Los genes en los módulos COL5A2 están mayormente sobreexpresados en todos los subtipos moleculares** Esta figura muestra la expresión de la firma de expresión de los genes que pertenecen a los módulos COL5A2 en **A)** Luminal A, **B)** Luminal B, textbfC) HER2+ y textbfD) subtipo basal de cáncer de mama. 180
5.11. **Estructura modular en los componentes 2 y 6 del subtipo basal. A)** Los colores definen cada módulo. **B)** Flujo de información entre las comunidades. El ancho del enlace es proporcional a la cantidad de enlaces compartidos entre los módulos. Las comunidades a rellenas en color representan aquellas que están enriquecidas con alguna categoría de *Gene Ontology* (GO). 181
5.12. **Procesos de *Gene Ontology* (GO) asociados con módulos en la red de transcripción del subtipo basal de cáncer de mama.** En esta figura, los módulos se colorean de acuerdo con el código de color de la Figura 5.11. Estas comunidades están conectadas a las categorías de GO que están coloreadas de acuerdo con un proceso general. 183
5.13. **Los procesos de muerte celular e infección viral están regulados de manera opuesta por la misma firma molecular del subtipo basal de cáncer de mama.** En esta figura, los genes se representan según sus niveles de expresión: rojo para sobreexpresado y azul para genes subexpresados. Las líneas entre las moléculas y procesos indican la función prevista de la molécula de acuerdo con su valor de expresión, la línea azul conduce a una inhibición del proceso; a su vez, las líneas naranjas representan la activación prevista. Los procesos de color de **muerte celular** y **infección viral** representan el mismo efecto predicho que las líneas. 184
5.14. **Componente gigante de la red del subtipp Her2**. **A)** Se muestra la arquitectura de red de Her2. **B)** Genes expresados diferencialmente, los nodos rojos representan genes sobreexpresados, mientras que los nodos azules son genes subexpresados. **C)** Estructura modular de la red de cáncer de mama Her2+ **D)** Estructura submodular. Nótese que en B), los genes con un patrón de expresión similar se agrupan. 186
5.15. **Procesos biológicos enriquecidos por módulo en la red del subtipo Her2+ de cáncer de mama.** La figura muestra los procesos enriquecidos (renglones a la derecha) por módulo (columnas con nombres en la parte inferior). Nótese que el módulo LCK tiene la mayoría de los procesos enriquecidos, pero los procesos no enriquecidos en LCK se enriquecen en los demás módulos. El gradiente de color es proporcional al $-log(p_v)$ de los procesos enriquecidos. 188

5.16. **Perfil de expresión del módulo COL5A2**. Este submódulo está compuesto principalmente por genes sobreexpresados. Interesantemente, los genes subexpresados (representados en azul claro) tienen un pequeño número de conexiones, en comparación con el número de enlaces que tienen la mayoría de los genes sobreexpresados. Nótese que los genes incluidos en este módulo son la Familia de Colágeno, además de FN1, FBN, VCAN, LUM y THBS2. 189
5.17. **Perfil de expresión del módulo CNR2**. En esta figura, es evidente el módulo sobreexpresado en la parte inferior; este módulo está enriquecido para procesos asociados a la membrana plasmática. 190
5.18. **Perfil de expresión del submódulo ZBTB38 del módulo CNR2**. Este módulo está sobreexpresado en la mayoría de sus genes. Nótese que los genes DICER y AGO3 (resaltados), que codifican las proteínas DICER y Argonnauta3, parecen estar sobreexpresados, lo que sugiere un papel importante de esta comunidad en la regulación gen-microARN. . . 191
5.19. **Perfil de expresión del módulo LCK y respuesta viral**. A) El perfil de expresión del módulo LCK se muestra aquí. Observe la clara separación entre los pocos genes subexpresados a la izquierda de la figura. B) El submódulo en el cual el gen OAS2 es el de mayor PageRank, también está enriquecido para varios procesos relacionados con la respuesta de virus. Esto está de acuerdo con los resultados obtenidos en la figura 5.20, donde se prevé que los procesos de infección viral se activan. 192
5.20. **Enfermedades y funciones enriquecidas en el subtipo Her2+ de cáncer de mama**. Esta figura muestra las enfermedades y funciones enriquecidas del subtipo molecular Her2+ de cáncer de mama, de acuerdo con la *Ingenuity Knowledge Base* (IKB). Los tamaños cuadrados son proporcionales al $-log(p_v)$ del conjunto de genes. El código de color representa el *z-score* de las moléculas en cada categoría, dependiendo de la actividad pronosticada que el proceso podría ejercer: los colores azules tienen en cuenta la disminución del proceso, mientras que los colores naranja muestran procesos activados. La mitad superior contiene todas las categorías enriquecidas, separadas por categorías generales. La mitad inferior es un zoom en las categorías "Enfermedades infecciosas" (cuadrado rojo en la mitad superior). Obsérvese como todos los procesos, excepto las enfermedades respiratorias (gris) y las infecciones bacterianas (azul) tienen una activación prevista, lo que significa que las categorías están más activadas que en un control. 193

Índice de tablas

5.1. **Principales parámetros topológicos de la red transcripcional MEF2C.** 167
5.2. **Parámetros de la estructura modular de subtipos moleculares de cáncer de mama.** 174
5.3. **Módulos enriquecidos encontrados en redes transcripcionales para subtipos moleculares de cáncer de mama.** Los módulos están etiquetados de acuerdo con sus genes con mayor PageRank (sección 2.2.5.1).177

Capítulo 1

Ciencia de Redes aplicada a la Biología.

Uno de los más interesantes desafíos de la biología hoy en día es dilucidar cómo una secuencia unidimensional de nucleótidos en el *Ácido desoxirribonucleico* (ADN o DNA en inglés) dirige el desarrollo de un organismo en un patrón de tipos de células (*fenotipo*), tanto en organismos procariontes como en eucariontes. Es una realidad que los estudios bioquímicos y genéticos han aumentado enormemente la comprensión de la *síntesis de proteínas*, sin embargo, se sabe poco acerca de cómo se coordina la ***regulación de genes*** individuales dentro de una célula (*regulación génica global*) para producir una célula diferenciada, y tampoco se entiende del todo, cómo se coordina la formación de células diferenciadas dentro de un organismo, a partir de la información genética.

Entender este intricado *programa de regulación*, podría ayudar a comprender mejor la diferenciación celular en el contexto de la *biología evolutiva del desarrollo* (evolutionary developmental biology)[1], así como en la biología que subyace detrás de enfermedades complejas como el cáncer. Asimismo descifrar el programa de regulación en el genoma humano podría en un futuro coadyuvar a vislumbrar posibles opciones y alternativas terapéuticas para estas enfermedades.

En el presente estudio presentaremos una forma de modelar el programa de regulación genético y entenderlo mediante el uso de Redes Complejas y sus propiedades topológicas de gran escala. Asimismo tiene como objetivo principal presentar una metodología computacional matemáticamente bien fundamentada para encontrar módulos (sub-unidades) funcionales dentro del programa de regulación genetica modelado por redes de *regulación genética de gran escala* (es decir, compuestas de miles de genes e interacciones transcripcionales).

[1] lo que hoy en día se conoce como *Evo-Devo*.

El programa de regulación transcripcional genético.

Partiendo de *Dogma Central de la Biología Molecular* [1, 2], un gen X, puede transcribirse en moléculas de *Ácido Ribonucleico* (ARN) *mensajero*, mediante la enzima *RNA-polimerasa*. Estás moleculas de ARN mensajero X_{mRNA} asociadas al gen X son posteriormente traducidas en proteínas X_P en la célula mediante *ribosomas*. Así. la tasa de síntesis de proteínas X_P a partir de un gen X es un complejo proceso que pude ser regulado de varias maneras que aún no conocemos totalmente.

La concentración presente (baja o alta) de ciertas proteínas, o bien la ***expresión*** *de los genes* asociados a dichas proteínas, es lo que hace diferentes a las células y les confiere entre otras cosas su *fenotipo* particular [3]. En efecto, el hecho de que dos células sean diferentes aunque tengan exactamente la misma información genética radica en que tengan dichos genes *expresados* de forma diferente. Dado lo anterior, regular la tasa de síntesis (transcripción→traducción) de proteínas en las células está directamente relacionado a la diferencia entre éstas. Así, la regulación de la síntesis de proteínas o bien ***regulación transcripcional genética***, es el fenómeno mediante el cuál la tasa de traducción de un gen X en proteína X_P es regulada mediante varios factores. Uno de estos puede ser que una molécula inhiba o promueva la transcripción y eventual traducción del gen X.

El proceso de Regulación Transcripcional Genética.

La cadena de bases nitrogenadas (con el alfabeto A,C,G,T)[1] que componen a un gen, se encuentra flanqueada por otras secuencias de bases, en particular "al principio" del gen, mejor dicho *rio arriba*, en el extremo $5'$ del gen, donde se encuentra una región *promotora* de la transcripción, compuesta de una secuencia que puede reconocer la *RNA-polimerasa* para unirse a esa región de ADN, y en seguida una región llamada *operador*, para promover la transcripción. En células eucariotas estas regiones se conocen como las llamadas *cajas TATA*, ya que son secuencias de Adenina y Timina en el ADN. A estás cajas TATA se puede unir una molécula (por lo general una proteína) *activadora* o bien *represora* de la transcripción llamada ***Factor de Transcripción*** (**TF**). Los *sitios de unión al ADN* de los Factores de Transcripción también suelen llamarse **TFBS** (*Transcription Factor Binding Sites*).

Así, cuando hay una ***regulación positiva***, una proteína **activadora** (*Factor de Transcripción*) se une al operador (*caja TATA* o **TFBS**), para reclutar a la *RNA-polimerasa* y así promover la transcripción. Por lo que en ausencia del *Factor de Transcripción* activador, no hay transcripción. Por el otro lado, en una ***regulación negativa***, un *Factor de Transcripción* **represor** se une al **TFBS** (secuencia TATA u operador) impidiendo que la *RNA-polimerasa* pueda unirse al ADN y así se evita la transcripción.

[1] **A** ≡ Adenina, **C** ≡ Citosina, **G** ≡ Guanina, **T** ≡ Timina.

Solo en ausencia de la proteína (**TF**) represora se puede transcribir el gen. La figura 1.1 muestra de forma esquematica el proceso de regulación negativa y positiva.

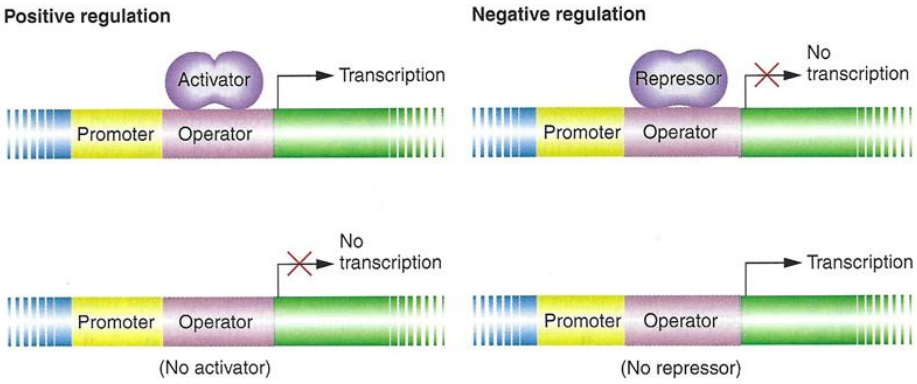


Figura 1.1: Esquema de regulación transcripcional *positiva* y *negativa*.

Dado que los Factores de Transcripción son en general proteínas, entonces fueron transcritos y traducidos a partir de algún otro gen en el ADN. De ésta manera un gen X (transcrito en un X_{mRNA}) que eventualmente se traduce en un factor de transcripción TF_X puede *promover* o *inhibir* la traducción de otros genes (A, B, C, D, F, ... etc.) cuyas proteínas pueden a su vez pueden ser Factores de Transcripción de transcripción de otros genes incluso del propio gen X. Lo anterior se ejemplifica en la figura 1.2.

Este proceso se puede representar mediante lo que se conoce matemáticamente como una *gráfica* (ver sección 2.1, en el capitulo 2), en la cual puntos (llamados *nodos* o vértices) que representan genes se unen mediante flechas o lineas (llamadas *aristas*) que representan una regulación positiva o negativa (figura 1.3). Como veremos más adelante (en el capítulo siguiente), estás gráficas están matemáticamente bien definidas y cuando son de muchos nodos (*Redes Complejas*) cuentan con propiedades estructurales y estadísticas que las hacen un excelente modelo de la regulación transcripcional.

Así, en muchos organismos existen millones de interacciones transcripcionales entre miles de especies químicas que son producto de la información genética (*e.g.* el genoma humano)[1]. A éstas *redes complejas* de genes e interacciones que regulan la transcripción de otros genes y/o de ellos mismos, la llamaremos *Programa de Regulación Transcripcional Genético*.

[1]El genoma humano cuanta con alrrededor de 22,000 genes. Muchos de estos codifican a proteinas que son factores de transcripción, incluso de varios genes, con lo cual se tiene una red de regulación muy grande.

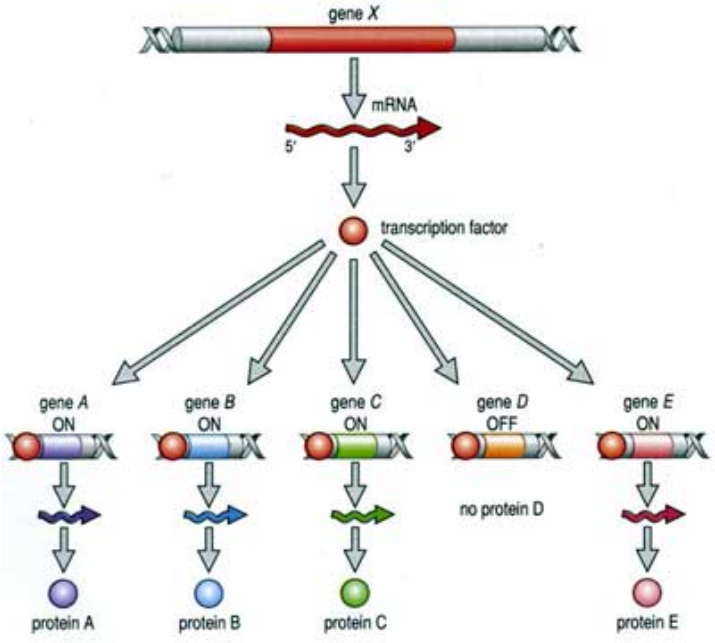


Figura 1.2: Esquema de funcionamiento de factores de transcripción.

Dado que la tasa de síntesis (transcripción→traducción) de una proteína está directamente asociada al fenotipo de la célula donde se lleva a cabo, regular estás interacciones es esencial para organismos procariontes, eucariontes e incluso virus. Ya que aumenta la versatilidad y adaptabilidad de un organismo al permitir que las células *expresen* proteínas cuando es necesario. En el caso humano, desajustes o errores en el *Programa de Regulación* podrán conllevar a fenotipos diferentes no siempre funcionales (*enfermos*). Por lo cuál modelar la regulación transcripcional del genoma completo en humanos mediante herramientas teoricas (matemáticas y/o computacionales) cobra mucha relevancia en el contexto de optimizar recursos experimentales con miras a una *medicina genómica*.

Por último, es importante señalar que por lo que se refiere a biología molecular, la modelación por redes no se limita al estudio de la regulación genética, también se aplican al estudio de interacciones de proteínas (*Protein-Protein Interactions* o PPI) (ver sección 3.1.2, capitulo 3) y también a relaciones entre moléculas en una reacción metabolica (Redes Metabolicas). Sin embargo el presente trabajo esta dedicado al estudio de Redes de Regulación Genética o *Gene Regulatory Networks* (GRN).

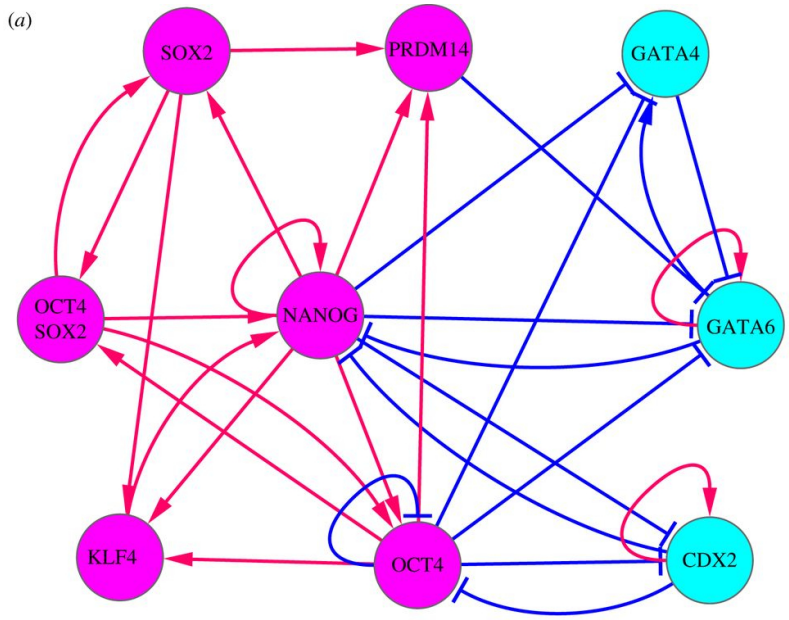


Figura 1.3: Modelación de la regulación de algunos genes mediante una gráfica pequeña. Las flechas rojas representan una *regulación positiva* (promoción) de la transcripción del gen de donde parten hacía donde llega la flecha. Las flechas azules representan una *regulación negativa* (o inhibición) de la transcripción. Cuando una flecha entra y sale del mismo gen (nodo) representa una autoregulación.

Redes de Regulación genética clásicas.

Las redes de regulación genética tienen su origen en la identificación del *operón lac* de *E. coli* reportado en los trabajos de Jacob y Monod de principios de los 1960s [4]. Los autores argumentaron que algunas enzimas implicadas en el metabolismo de la *lactosa* son expresadas por *E. coli* sólo en presencia de lactosa y ausencia de glucosa. El descubrimiento de este sistema de regulación de genes ha sido considerado ampliamente en la literatura y dio pie al planteamiento de los primeros trabajos para modelar el proceso de regulación genética en bacterias, mediante redes (pequeñas). De esta manera los trabajos propuestos por E.H. Davidson[1] y S. A. Kauffman han sido de los más influyentes respecto a modelar la regulación global coordinada de genes dentro de una célula.

En 1969, R.J. Britten y Davidson [6, 7], publicaron el primer modelo de una red de regulación genética. Esta incluía tanto secuencias de ADN reguladoras, es decir, seg-

[1]E.H. Davidson falleció recientemente en 2015, su trayectoria y aportación a las Rede de Regulación Genetica es invaluable [5]

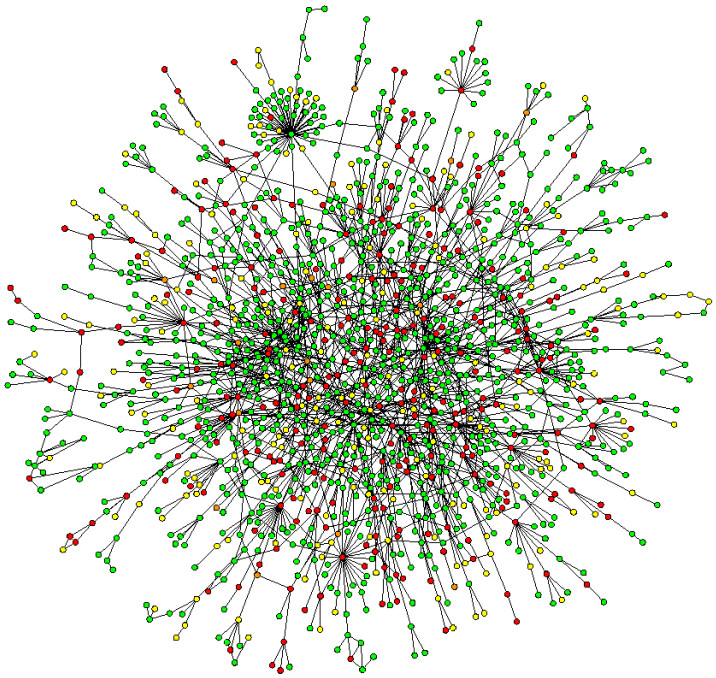


Figura 1.4: Red de interaccción de proteinas en Sacaromise cerevisiae.

mentos de ADN en los que cada gen especificaba las condiciones de donde podría ser "encendido" (transcrito), y también genes que codifican moléculas que se unen al ADN regulador de otros genes para "encenderlos" o "apagarlos" (factores de transcripción). Los autores también prevén una **jerarquía** de genes en la que la expresión de *genes estructurales* está controlada por relativamente pocos complejos de genes integradores. Así también, señales positivas como hormonas, determinan qué integradores se expresan en cada célula. La teoría expuesta no sólo hacia referencia a organismos procariontes, sino que trataba de expandir las ideas planteadas a células y organismos más complejos.

El trabajo de Davidson y colaboradores se amplió a otros organismos (figura 1.5), destacando la descripción de la red de regulación genética del *erizo de mar* [8, 9]. Su grupo realizó un mapeo sistemático de pruebas funcionales para detectar todas las conexiones de control entre los genes involucrados en los eventos clave en las primeras etapas del desarrollo del embrión de erizo de mar, definiendo casi todas las redes de genes que especifican el desarrollo en los cinco tipos de tejidos del erizo de mar de 30 horas de edad, desde la fertilización del huevo hasta el desarrollo de un organismo [10, 11]. Así, los autores reconocieron que las redes de regulación que rigen los procesos de alto nivel (la formación de un fenotipo específico de célula, por ejemplo) se construyen a partir de circuitos de genes que pueden tener sorprendentes similitudes incluso cuando las identidades de los genes en el los circuitos en sí son diferentes y publicaron el primer modelo

computacional completo de la red de embriones de erizo de mar, que consta de unos 50 genes reguladores clave [12].

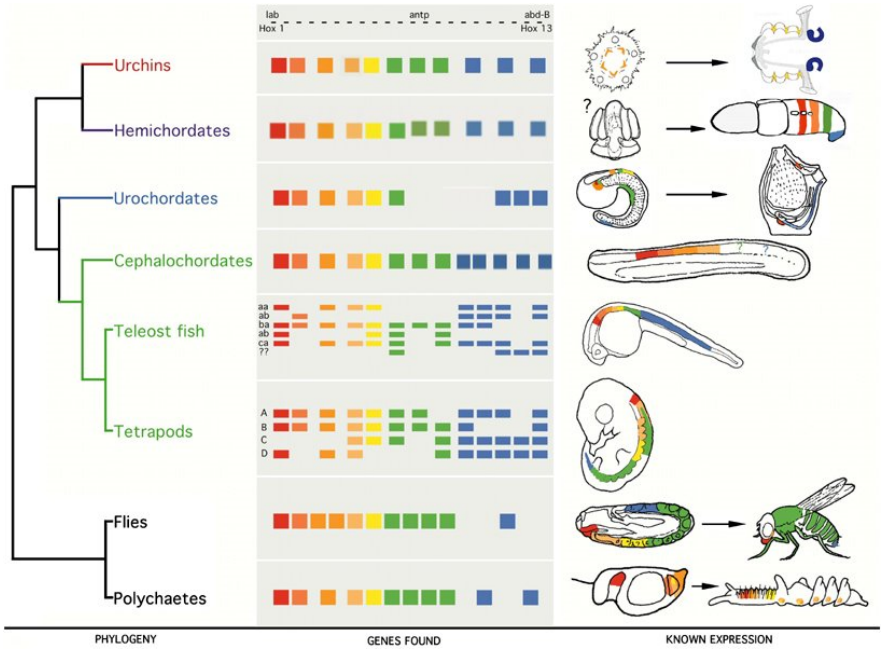


Figura 1.5: Comparación de algunos genes involucrados redes de regulación en el desarrollo corporal en diferentes organismos.

Además Britten y Davidson fueron pioneros en describir las redes de regulación en el desarrollo corporal y evolutivo de animales [7], las cuales se componen de diversos elementos que evolucionan a diferentes velocidades y de diferentes maneras. Debido a la organización jerárquica de las redes de regulación en el desarrollo, algunos tipos de cambios afectan las propiedades terminales del plan corporal, como ocurre en la especiación, mientras que otros afectan los aspectos principales de la morfología [13] (figura 1.6). Estos circuitos lógicos [14] se pueden ver como unas pocas docenas de *tipos de módulos* que realizan funciones específicas (ver sección 3.1.2, capitulo 3), como crear un ciclo de retroalimentación positiva que estabilice la expresión de un par de genes en un futuro tipo de tejido, o un circuito de exclusión que cree barreras entre diferentes tipos de tejido en algunas partes del futuro organismo.

Así, el trabajo de Davidson *et al.* [15, 16] abrió el camino a las demostraciones de que sistemas modulares similares parecen existir en otros organismos como moscas, ranas, los embriones de pollo, ratones y el peces, lo que sugiere que tales módulos lógicos pueden ser una característica universal de los organismos superiores [11, 17]. Davidson señaló en su momento, que su trabajo permitiría a los científicos manipular y rediseñar

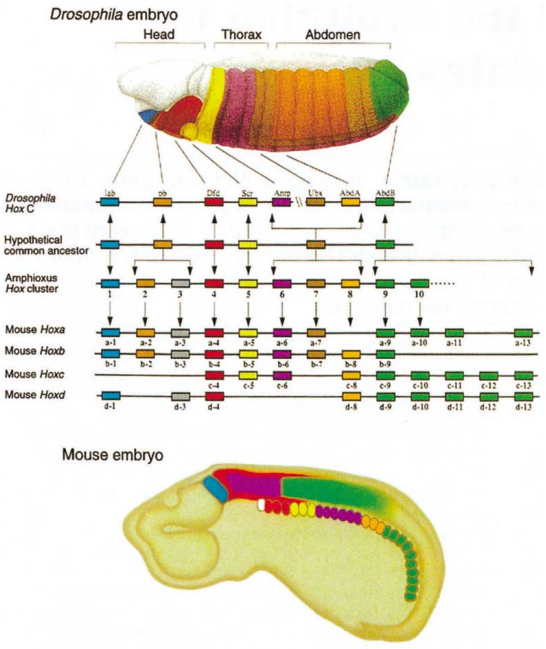


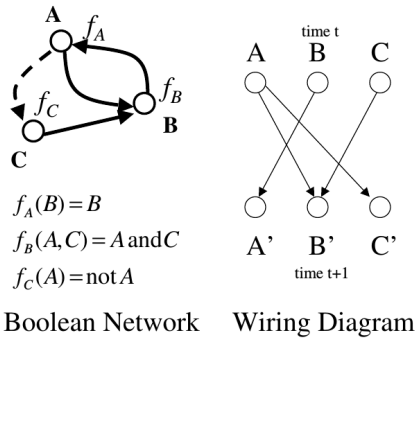
Figura 1.6: Comparación de módulos en redes de regulación genética asociadas al desarrollo en mosca (*Drosophila melanogaster*) y ratón (*Mus musculus)* .

redes genéticas, como un proceso que simularía los cambios genéticos que acompañan la evolución y el desarrollo de los organismos en la vida real [18, 19], en sus propias palabras: "La evolución de los animales se debe a los cambios en la estructura de estas redes reguladoras de genes, por lo que este trabajo nos brinda la oportunidad de estudiar la evolución de una manera nueva y decisiva" [20].

De manera paralela al trabajo de Davidson *et al.*, las ideas de Jacob y Monod fueron incorporadas en un marco matemático por varios investigadores, incluyendo Sugita [21], Thomas [22] y Stuart A. Kauffman [23, 24]. La idea principal de estos primeros trabajos pioneros era representar a los elementos de regulación genética como si fuesen circuitos lógicos similares a los que se encuentran en los microprocesadores de las computadoras. Así, es posible modelar la regulación de los genes como *Circuitos Booleanos* con una dinámica temporal discreta, donde cada gen cuenta con un estado también discreto en el que está "sintetizándose en proteína" o no (*activo* o *inactivo*), 1 o 0, en cada tiempo.

El más influyente de estos modelos es el de S. A. Kauffman [23, 24]., en el cual al tiempo t cada gen representado por un nodo, está en un estado prendido 1 (*expresándose*) o apagado 0 (*reprimido*). Además cada nodo tiene k vecinos que son los genes que lo regulan, los cuales a su ves están en cualquiera de los dos estados. El estado de los genes en el siguiente tiempo $t + 1$ está dado a partir de función (o tabla de verdad

lógica) (figura 1.7) construida a partir de operadores lógicos booleanos AND ($\wedge$), OR ($\vee$) y NO ($\neg$) cuyas entradas son los estados de los genes en el tiempo anterior t. Esta tabla de verdad es equivalente a una función de transición (de los nodos a su siguiente estado) de un *autómata finito determinista* en *teoría de la computación*. De hecho, los autómatas celulares booleanos [25], son casos particulares de las redes booleanas, donde se determina el estado de una variable por sus vecinos espaciales.



State	INPUT			OUTPUT		
	A	B	C	A'	B'	C'
1	0	0	0	0	0	1
2	0	0	1	0	0	1
3	0	1	0	1	0	1
4	0	1	1	1	0	1
5	1	0	0	0	0	0
6	1	0	1	0	1	0
7	1	1	0	1	0	0
8	1	1	1	1	1	0

Truth Table

Figura 1.7: Tabla de verdad: función de actualización de una red booleana.

Así una red de Kauffman, se compone de una gráfica $G(V, E)$ (ver sección 2.1 del siguiente capítulo)[1], anotada con un conjunto de estados $X = \{x_i \mid i = 1, \ldots, n\}$, junto con un conjunto de funciones booleanas $B = \{b_i \mid i = 1, \ldots, k\}$, donde cada nodo i, tiene asociada una función de transición a partir de los estados de los nodos conectados a i. De esta manera los estados de cada nodo i en la red al tiempo $t+1$ viene dado por:

$$x_i(t+1) = b_i(x_{i1}, x_{i2}, \ldots, x_{ik}) \tag{1.1}$$

Donde $x_i(t)$ es el estado del nodo i en el momento t y x_{ij} son los estados de los nodos conectados a i.

De esta manera se modela una Red de Regulación Genética de n genes en estado binario con k entradas en cada nodo que representa los mecanismos de regulación. Los dos estados (encendido / apagado) representan, respectivamente, la condición que un gen se activa o inactiva (figura 1.8). Así, el estado de toda la red en cualquier tiempo está dado por los estados actuales de todos los n genes, por lo que el espacio de estados de cualquier red, es 2^n, dado que cada gen puede estar en dos estados (0 y 1). De esta manera, tarde o temprano llegará a un estado visitado anteriormente y, por lo tanto, ya que la dinámica es determinista, caerá en un estado *atractor*. Si el atractor se compone de más de un estado, se denomina un *ciclo atractor*. El conjunto de estados que

[1] En el siguiente cápitulo se expondrá ampliamente las definiciones matemáticas de las redes.

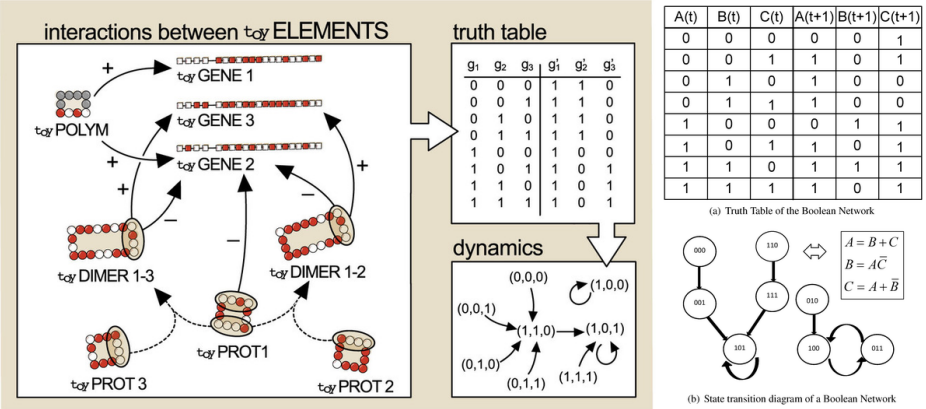


Figura 1.8: Ejemplos de aplicación del modelo de Kauffman.

conducen a un atractor se llama *cuenca de atracción*. Los estados sin predecesores se llaman estados *jardín del Edén* y la dinámica de la red fluye desde estos estados hacia los atractores (figura 1.9).

Kauffman *et al.* [23, 24, 26] demostraron que estos modelos, en los que los genes controlan no más de dos o tres otros genes, pero que contienen señales de realimentación, pueden contener información suficiente para producir la regulación autónoma de muchos genes dentro de una sola célula [27] . Además Kauffman, propuso que *atractores* en esta red discreta correspondían a tipos de células en los organismos, entre otras propiedades biológicas [28]. Así, el trabajo de Kauffman sobre el análisis matemático de las propiedades de las redes booleanas, a menudo llamadas *redes de Kauffman*, ha sido un área de intenso interés teórico [29, 30, 31, 32, 33].

De esta manera a partir de los trabajos pioneros de Davidson y Kauffman el trabajo sobre redes de regulación genética se desarrollado para describir organismos cada vez más complejos. Como por ejemplo en la levadura (*Saccharomyces cerevisiae*), donde se determinado que la mayoría de sus reguladores transcripcionales codificados se asocian con genes a lo largo del genoma en células vivas, lo cual revela que las funciones celulares en eucariotas están altamente conectadas a través de redes de regulación transcripcionales que regulan otros reguladores transcripcionales [34].

Así también, el trabajo en erizo de mar [35, 36, 37] y la mosca (*Drosophila melanogaster*) [38, 39] ha demostrado que se pueden generar diagramas de red de regulación transcripcional razonables que representan el desarrollo temprano en animales multicelulares mediante el uso de herramientas genómicas, genéticas y bioquímicas apropiadas. Lo anterior, abre la posibilidad para el establecimiento de redes reguladoras similares para el desarrollo de vertebrados. [40]. Otro ejemplo es de la morfogénesis de flores en *Arabidopsis thaliana*, donde se han derivado restricciones de parámetros que representan

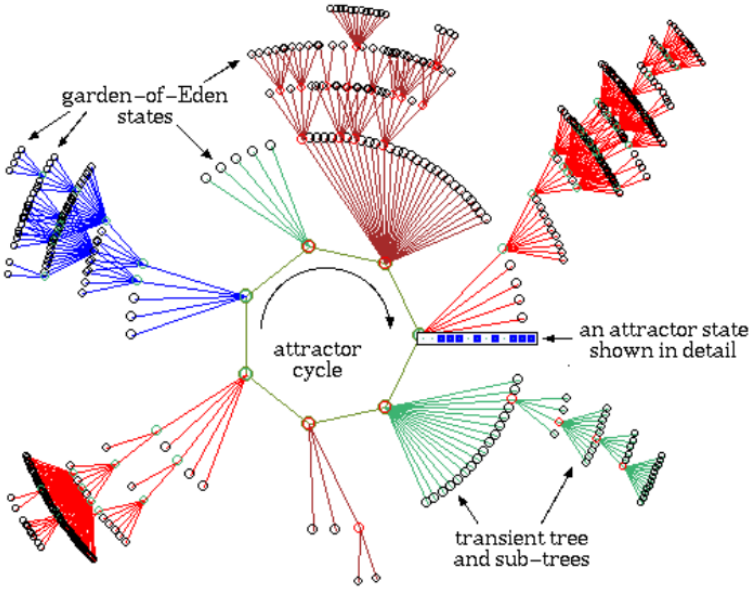


Figura 1.9: Ejemplos de aplicación del modelo de Kauffman.

los diferentes patrones de expresión génica encontrados en los cuatro órganos florales (sépalos, pétalos, estambres y carpelos), además de un estado "no floral", lo que ha llevado a proponer algunas conclusiones generales sobre la estructura de las redes de genes que controlan el desarrollo en plantas [41, 42] (figura 1.10).

De esta manera las redes de regulación genética se han convertido en un marco general de trabajo en el estudio de la *biología evolutiva del desarrollo* [43, 44, 45, 46, 47], no sólo en organismo sino también el estudio de células humanas como las del sistema inmune, por ejemplo, la diferenciación y especialización de *células B* [48]. Y asimismo los métodos de GRN se han refinado, ampliado y complementado mediante otros enfoques como las *Ecuaciones Diferenciales* (ordinarias y estocásticas) [27, 49, 50, 51] o bien las redes boolenas probabilisticas y estocásticas [52, 53, 54].

Sin embargo, la desventaja principal del enfoque de este tipo de Redes de Regulación es que para genomas completos muy grandes es difícil inferir una red total de regulación a partir de los datos experimentales, los cuales muchas veces, no son suficientes. Además de que no explican cómo los patrones de células diferenciadas se forman de forma autónoma en la regulación genética global. Asimismo en el caso de las redes boolenas tratar de modelar una red transcripcional para genoma completo en humano es computacionalmente inviable pues el numero de estados posibles sería enorme así como inferir todas las reglas de actualización es muy difícil. Aún así, este enfoque destaca que las redes son modulares y jerárquicas en etapas de diferenciación o desarrollo celular, por

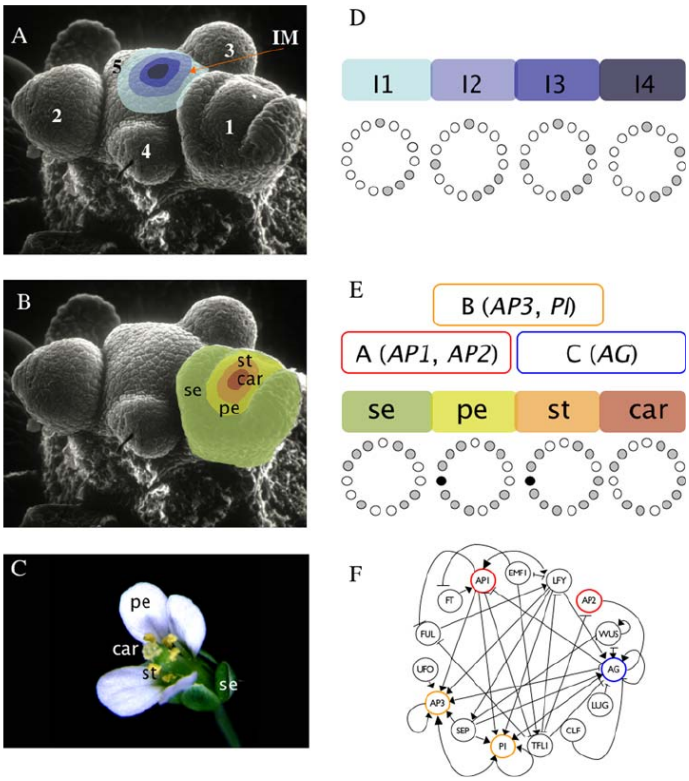


Figura 1.10: Modulos en la red de regulación genética subyacentes a la determinación de la célula-destino del órgano floral primordial asociada al desarrollo de flores en *Arabidopsis thaliana*. Figura tomda de Alvarez-Buylla *et al.*, [42]

lo que contar con enfoque modular en la modelación de la regulación genética cobra muchas relevancia.

Bioinformática y Biología de Sistemas.

Hasta ahora hemos expuesto el enfoque clásico de las Redes de Regulación Genética, sin embargo en las últimas décadas, los rápidos desarrollos en tecnologías de investigación genómica y los avances en las tecnologías de la información se han combinado para producir una gran cantidad de información relacionada con la biología molecular.

A partir del desarrollo de estas técnicas experimentales hemos sido testigos de un cambio de enfoque en las ciencias biomoleculares, desde el deseo de comprender cómo funciona un solo gen hasta comprender cómo todos los genes y productos genéticos de

una célula funcionan en conjunto. Así, con el advenimiento de la era de la información y sus nuevas tecnologías, grandes volúmenes de datos metabólomicos, proteómicos y genómicos se almacenan todos los días en grandes bases de datos publicas. Lo anterior ha propiciado propuestas de nuevos enfoques para inferir Redes de Regulación a partir de de estos datos. Con lo que tenemos la oportunidad de poder modelar genomas completos de organismos más complejos como el humano y aplicarlo a enfermedades complejas como el cáncer.

Herramientas para el análisis de datos genómicos.

A la par del desarrollo de técnicas experimentales como los métodos para determinar la secuencia de trozos cortos de ADN , producir más ADN en una célula bacteriana, clonar un gen, amplificar secuencias de ADN in vitro a través de la reacción en cadena de la polimerasa (PCR), etc., la ciencias de la computación se han vuelto esenciales para la Biología Molecular. De está manera, nació lo que hoy conocemos como ***Bioinformática*** que es un campo interdisciplinario que desarrolla métodos y herramientas de software para comprender datos biológicos y la cual combina informática, biología, matemáticas y aplica estadística para analizar e interpretar datos biológicos.

El campo de la *bioinformática* experimentó un crecimiento explosivo a partir de finales de la década de 1990, impulsado en gran medida por el *Proyecto del Genoma Humano* [55] y por los rápidos avances en la tecnología de secuenciación del ADN. De está manera ha evolucionado en el análisis e interpretación de varios tipos de datos. Esto incluye secuencias de nucleótidos y aminoácidos, dominios de proteínas y estructuras de proteínas. Asimismo, los usos comunes de la bioinformática incluyen la identificación de genes candidatos y polimorfismos de un solo nucleótido (SNP), las cuales a menudo se realizan con el objetivo de comprender mejor la base genética de una enfermedad, adaptaciones únicas o diferencias entre las poblaciones.

Asimismo, la bioinformática implica la creación y avance de bases de datos, algoritmos, técnicas computacionales y estadísticas para resolver problemas formales y prácticos derivados de la gestión y el análisis de datos biológicos. Uno de los grandes ejemplos entre desarrollo experimental de la biología molecualr y nanalisis bioiformatico que permiten estudiar datos de genoma completo son son los estos avanaces son los microarreglos de expresión

Un microarreglo de expresión de (o chip de ADN) es una superficie sólida (vidrio o plástico), a la cual se une una colección de fragmentos de ADN (figura 1.11). Se usa para analizar y monitorear de manera simultánea los niveles expresión de miles de genes. **El nivel de expresión de un gen es la cantidad de ARN mensajero que produce**.

Así el funcionamiento de los microarreglos consiste, básicamente, en medir el nivel de hibridación entre una sonda específica (probe), y una molécula diana (target). Estos

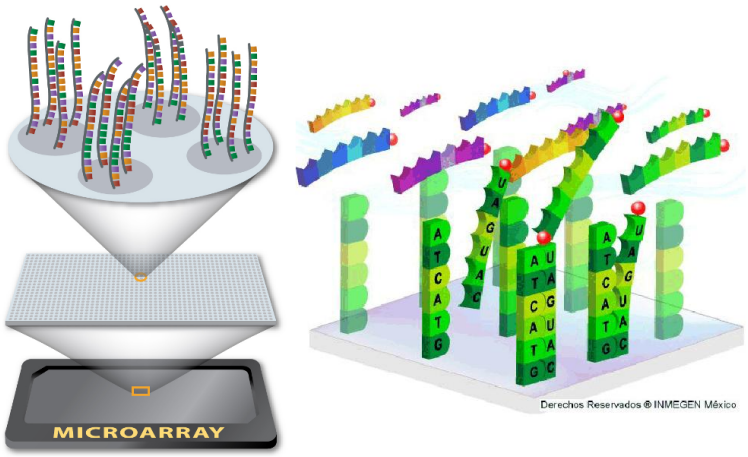


Figura 1.11: Microarreglo de expersión.

targets son secuencias de ARN mensajero producidos por la células de las que se toma la muestra. Asimismo, los probes tienen adherida una molécula fluorescente que se activa cuando el el target se une al probe. Así, es posible medir el nivel de expresión de un gen mediante fluorescencia, a través de un análisis de imagen, que indica el nivel de expresión del gen.

Finalmente, hay que remarcar que el objetivo principal de la bioinformática es aumentar la comprensión de los procesos biológicos, por lo continuamente desarrolla e implementa programas informáticos que permiten el acceso, uso y gestión eficientes de varios tipos de información. Sin embargo, la bioinformática, por si sola, no genera un nuevo marco teórico para el estudio de los sistemas biológicos capaz de plantear hipótesis, sino que es parte de un nuevo enfoque integrativo conocido hoy día como *Biología de Sistemas*.

Biología de Sistemas.

Como mencionamos anteriormente, dada la enorme cantidad de datos que se generan y almacenan a partir de las ciencias biomoleculares, tenemos la posibilidad de cambiar de enfoque. En lugar de centrarse en los genes individuales y sus efectos, podemos atacar una pregunta simple, pero que requiere una gran cantidad de información para ser respondida: ¿cómo es la combinación de todos genes dentro de una célula son capaces de gobernar todas las reacciones que ocurren en la misma? Así, a partir de la basta cantidad de información y la avance en la modelación matemática de sistemas complejos, recientemente ha surgido un nuevo enfoque conocido como *Biología de Sistemas*, que pretende atacar preguntas como la anterior a partir técnicas matemáticas y computacionales. El Proyecto del Genoma Humano es un ejemplo del pensamiento de

sistemas complejos aplicado a la biología, que ha llevado a nuevas formas colaborativas de trabajar en el campo de la genética.

Así, la biología de sistemas puede definirse como el modelado computacional y matemático de sistemas biológicos complejos. Es un campo de estudio interdisciplinario basado en la biología que se centra en las interacciones entre los componentes de los sistemas biológicos, y cómo estas dan lugar a la función y comportamiento de ese sistema. Utiliza un enfoque holístico (en lugar del reduccionismo tradicional) para descifrar la complejidad de los sistemas biológicos que parte de la idea de que los organismos vivos son más que la suma de sus partes. Y asimismo, es colaborativo, integrando muchas disciplinas científicas (biología, informática, ingeniería, bioinformática, física y otras). Uno de los objetivos de la biología de sistemas es modelar y descubrir propiedades emergentes, en las células, tejidos y organismos (por ejemplo, las enzimas y metabolitos en una vía metabólica o los latidos del corazón), así como estudiar el comportamiento y organización de los procesos biológicos en términos de los componentes moleculares.

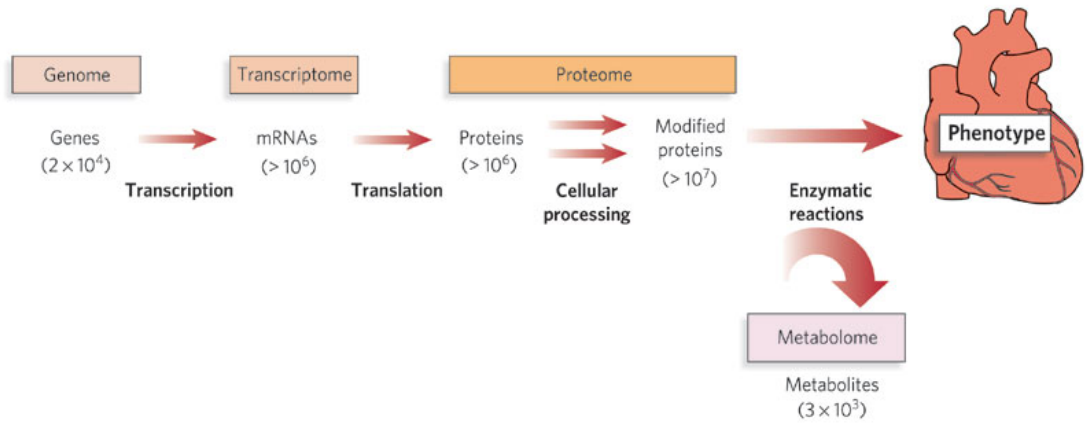


Figura 1.12: Estudio en diferentes "escalas *omicas*" de interacciones entre moléculas.

Así entonces, la capacidad de obtener, integrar y analizar conjuntos de datos complejos de múltiples fuentes experimentales utilizando herramientas interdisciplinarias, es parte central en la *Biología de Sistemas*; y algunos de sus campos de información más relevantes son: la Genómica, Epigenómica[1], Transcriptómica[2], Proteómica[3], meta-

[1]factores reguladores transcriptómicos no codificados en la secuencia genómica como metilación de ADN, acetilación de histonas y desacetilación, etc.

[2]Mediciones de expresión génica de tejido o célula completa mediante microarreglos de ADN o análisis seriados de expresión génica

[3]mediciones de proteínas y péptidos a nivel de organismo, tisular o celular mediante electroforesis en

bolomica, interactomica [1], entre otras (figura 1.12).

Sin embargo, no sólo la información es parte de la *Biología de Sistemas*, la construcción de un marco teórico para poder no sólo analizar sino generar modelos a partir de los datos es una parte fundamental. Esto implica el desarrollo de modelos mecanísticos, como los de los sistemas dinámicos a partir de las propiedades cuantitativas de bloques de construcción básicos. De hecho, el enfoque en la dinámica de los sistemas estudiados es la principal diferencia conceptual entre la *biología de sistemas* y la *bioinformática*. Por ejemplo, una red celular se puede modelar matemáticamente usando métodos que provienen de la cinética química y la teoría de control; utilizando conjuntos de ecuaciones diferenciales. Sin embargo, debido a la gran cantidad de parámetros, variables y restricciones en las redes celulares, a menudo se utilizan técnicas numéricas y computacionales para estos modelos. Esto hace que la *biología de sistemas* sea un tipo de ciencia difícil. Para muchos biólogos (y en particular biólogos moleculares), las ecuaciones eran algo que pertenecía a los laboratorios de física, y las matemáticas avanzadas y las estadísticas en profundidad eran una herramienta menos prevalente y opcional.

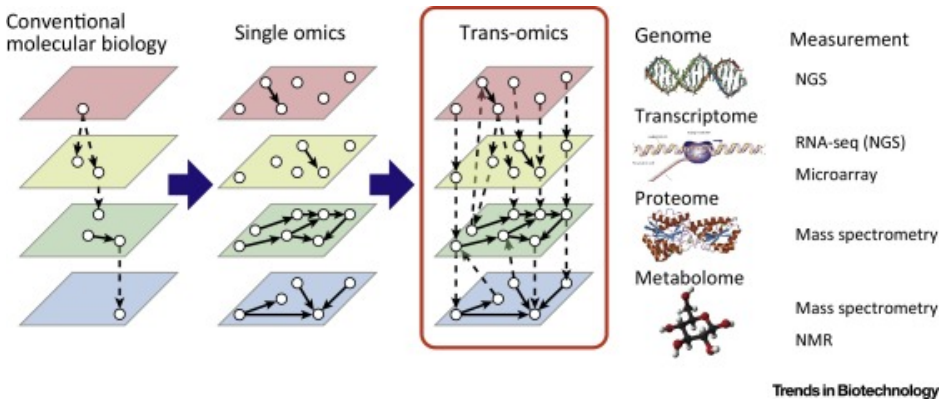


Figura 1.13: La *Biología de Sistemas* estudia a los sistemas biológicos en diferentes escalas, a partir de las grandes cantidades de datos *omicos* disponibles hoy en día.

Así los biologos de sistemas intentan resumir los datos acumulados en modelos matemáticos robustos. Cuyas recompensas pueden ser enormes, y su magnitud probablemente esté comenzando a emerger. En este sentido, la *biología de sistemas* se diferencia de los intentos anteriores de elaborar una serie de modelos físicos y matemáticos simples sobre lo que se llama "autoorganización biológica", dado que tiene los datos para validar o construir el modelo. Una amplia variedad de científicos cuantitativos (biólo-

gel bidimensional, espectrometría de masas o técnicas de identificación de proteínas multidimensionales (sistemas avanzados de HPLC acoplados con espectrometría de masas)

[1] Estudio a nivel de organismo, tisular o celular de interacciones entre moléculas.

gos computacionales, estadísticos, matemáticos, informáticos y físicos) están trabajando para mejorar la calidad de estos enfoques y para crear, refinar y volver a probar los modelos para reflejar con precisión las observaciones.

Asimsimo, la *biología de sistemas* comparte varios de los principios ideas fundamentales de los Sistemas Complejos. Como por ejemplo, se basa en el entendimiento de que el todo es mayor que la suma de las partes e intenta ofrecer una perspectiva más integrada sobre el funcionamiento interno de una célula.

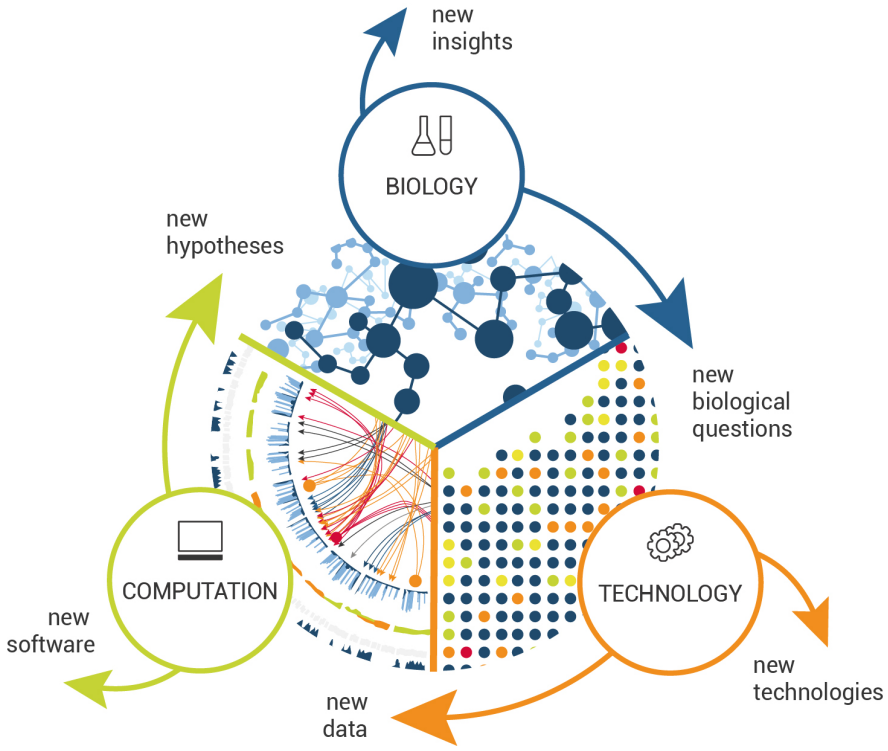


Figura 1.14: Esquema representativo que muestra el caracter integrativo, interdiciplinario y de retroalimentación de la *Biología de Sistemas*.

De está manera podemos entender a la *Biología de Sistemas* como una serie de protocolos operacionales utilizados para realizar investigación, es decir, un ciclo compuesto por modelos teóricos, analíticos o computacionales para proponer hipótesis comprobables específicas sobre un sistema biológico, validación experimental y luego usar la descripción cuantitativa recientemente adquirida de células o procesos celulares para refinar el modelo o teoría computacional (figura 1.14). Como el objetivo es un modelo de las interacciones en un sistema, las técnicas experimentales que más se adaptan a la *biología de sistemas* son aquellas que abarcan todo el sistema e intentan ser lo más

completas posible. Por lo tanto, transcriptómica, metabolómica, proteómica y técnicas de alto rendimiento se utilizan para recopilar datos cuantitativos para la construcción y validación de modelos (figura 1.13).

Inferencia de Redes de coexpresión de genes.

A partir de la información proveniente de los microarreglos de expresión genómica, se han propuesto nuevos enfoques estadísticos que permiten inferir redes de regulación genética, a menudo llamadas *redes de coexpresión de genes*, que exploran la funcionalidad de los genes a nivel de sistema. La inferencia de éstas redes de regulación genética se logra a parir de datos experimentales y se ha utilizado ampliamente para revelar interacciones entre genes a partir de sus niveles de expresión medidos experimentalmente.

La construcción de estás *redes de coexpresión de genes* se realiza a parir del nivel de expresión y es conceptualmente simple: los genes se conectan por parejas si se coexpresan significativamente a través de muestras de tejido elegidas apropiadamente [56, 57, 58, 59]. Uno de las métodos más empleados para calcular las correlaciones entre pares de genes y construir estas redes es el enfoque propuesto por Margolin *et al.* ARACNE [60].

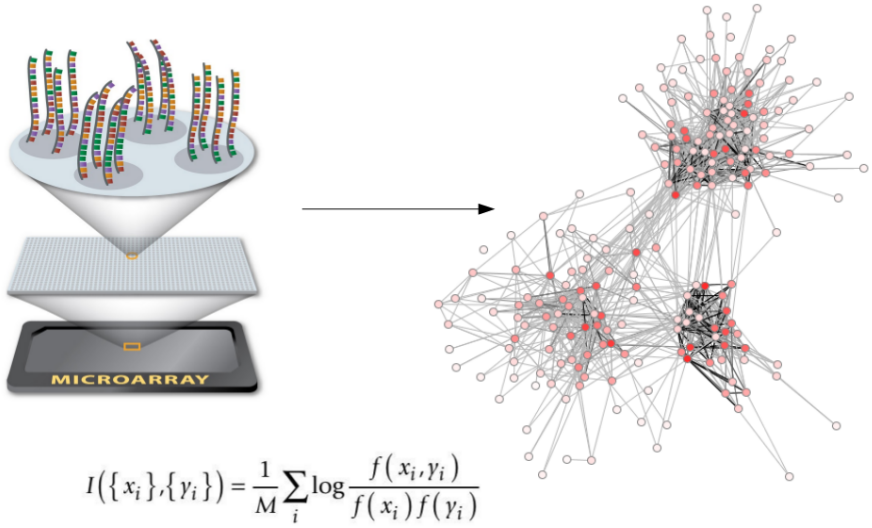


Figura 1.15: Inferencia de Redes transcripcionales de *co-expersión* a partir de *microarreglos*.

ARACNE, es un algoritmo, que utiliza perfiles de expresión de microarreglos, diseñado específicamente para escalar hasta la complejidad de las redes de regulación en

células de mamíferos. Este método utiliza como medida de *correlación* entre los niveles de expresión de pares de genes la ***Información Mutua*** [61, 62, 63, 64, 65, 66] entre ellos (ecuación (1.2). De está manera ARACNE captura correlaciones no lineales entre los niveles de expresión de genes sin importar que sean altos o bajos ya que utiliza la distribución de frecuencias completa para calcular la medida de información a diferencia de otras medidas de correlación como *Sperman* o *Pearson*.

$$I(X,Y) = I(\{x_i\},\{y_i\}) = \frac{1}{M}\sum_i log\frac{P(x_i,y_i)}{P(x_i)P(y_i)} \tag{1.2}$$

Donde x_i y y_i son los niveles de expresión de los genes X y Y, M es la tamaño de la muestra del microarreglo[1].

Así, ARACNE infiere redes en las que las aristas modelan posibles regulaciones transcripsionales directamente relacionadas a la variación conjunta del nivel de expresión de los genes incluso en células de mamíferos. Lo anterior nos provee de un modelo robusto con el cual contamos con todas las posibles interacciones transcripcionales para un genoma completo en humano (figura 1.15).

Las *redes de coexpresión de genes* han sido un tema ampliamente explorado en la literatura, en las que se han señalado sus fortalezas y debilidades [67] y asimismo de han utilizado exitosamente para construir *redes de coexpresión* en cáncer de mama [68, 69]. Es de destacarse que la inferencia de redes a partir de datos de coexpersión resulta en redes pesadas las cuales deben tratarse de forma diferente a sus contrapartes con enlaces no ponderados (véase sección 2.1.1; capitulo 2), y diseñar estrategias para atacarlas [58, 70, 71].

Asimismo, éstas redes son muy diferentes a las *Redes de Regulación booleanas* (clásicas) que hemos expuesto anteriormente, tanto en tamaño y dimensión. Las primeras derivadas de los enfoques de Kauffman y Davidson, son pequeñas y en general de organismos pequeños como levadura y bacterias. Mientras que estás últimas se infieren a partir de los datos de muchos experimentos de expresión contenidos en grandes bases de datos y que bajo el enfoque de Biología de Sistemas, es posible inferir redes de genoma completo. Asimismo estás redes al ser muy grandes cumplen las propiedades topológicas de las *Redes Complejas* que expondremos en el capítulo siguiente.

Redes *Complejas* aplicadas a la biología y enfermedades.

Como se expondrá ampliamente en el capítulo siguiente, en los últimos años nuestra comprensión de las redes ha experimentado una revolución, debido a la aparición de

[1] Se profundizara sobre la ***Información Mutua*** y *Teoría de la Información* en la sección 3.4.3.2 del capítulo 3 y en la sección 4.1.1.2 del capítulo 4.

una nueva serie de herramientas teóricas y técnicas para analizar redes reales. Estos avances han incluido algunas sorpresas que indican que la mayoría de las redes reales en sistemas tecnológicos, sociales y biológicos tienen diseños comunes que se rigen por principios organizadores simples y cuantificables. Así entonces es posible combinar los avances en el estudio de *Redes Complejas* con el enfoque de Biología de Sistemas para poder alcanzar una mejor comprensión de los sistemas biológicos. Lo anterior abre nuevas perspectivas en la medicina genómica y el tratamiento d enfermedades pues a partir de estás herramientas es posible modelar de forma teórica la interacciones (transcripsionales, entre otras) del genoma humano.

Así cuando se construyen redes muy grandes a partir de datos biológicos, es posible analizar matemáticamente sus propiedades topológicas dentro de un marco formal con el objetivo de revelar propiedades biológicas intrínsecas al sistema que la red representa. De está manera, la incursión de las *redes complejas* (o *ciencia de redes*) al estudio de sistemas biológicos bajo el enfoque de Biología de Sistemas ha dado paso a conceptos surgidos recientemente como el de **Biología de Red** (*Network Biology*) y **Medicina de Red** (*Network Medicine*).

Biología de Red (*Network Biology*).

Dado que la mayoría de los componentes celulares están conectados entre sí a través de intrincadas interacciones regulatorias, metabólicas y de proteína-proteína; el análisis de *redes complejas* está preparado para jugar un papel muy importante en la biología hoy en día. Por ejemplo, en una célula o microorganismo, los procesos que generan masa, energía, transferencia de información y especificación del destino de la célula se pueden integran a la perfección a través de una red (figura 1.16). Tomando el enfoque de biología de sistemas y la información generada a través de los experimentos en biología molecular, es posible generar una red de componentes y reacciones celulares, que a diferencia de una red pequeña sólo de regulación, es muy grande y tiene las propiedades de una *red compleja* (ver capítulo siguiente).

Para vislumbrar el papel clave de estas redes en el mantenimiento de las funciones celulares, Jeong *et al.* [72] analizaron matemáticamente la estructura a gran escala de redes metabólicas de 43 organismos que representan tres dominios de la vida y mostraron que, a pesar de la variación significativa en sus componentes y vías individuales, estas redes metabólicas tienen las mismas propiedades de escalamiento topológico además de mostrar sorprendentes similitudes con la organización inherente de sistemas complejos *no biológicos*. Esto puede indicar que la organización metabólica no solo es idéntica para todos los organismos vivos, sino que también cumple con los principios de diseño de redes robustas *libres de escala* (ver capítulo siguiente), y puede representar un modelo común para la organización a gran escala de interacciones entre todos los constituyentes celulares.

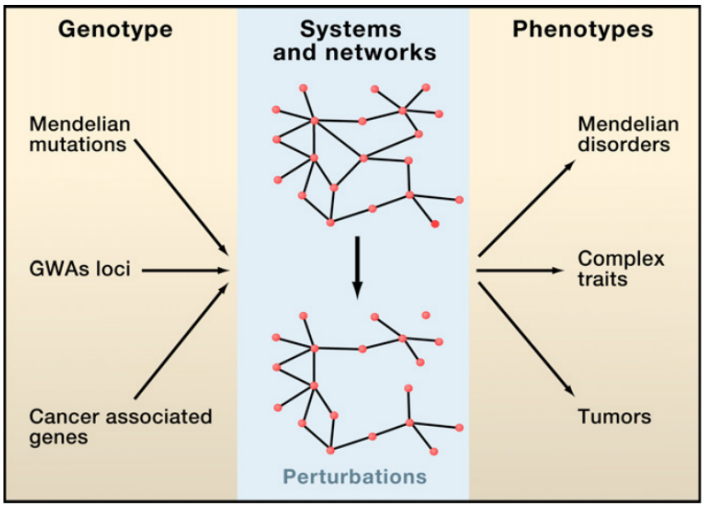


Figura 1.16: Esquema de como las redes celulares pueden ayudar a entender las relaciones genotipo-fenotipo. Las perturbaciones en los sistemas biológicos pueden modelarse como cambios estructurales en las redes complejas celulares que forma los genes y sus productos.

En general las redes biológicas se organizan ***modular*** y ***jerarquicamente*** (como vimos en las redes de regulación genética). En estás existen módulos funcionales (ver sección 3.1.2, capitulo 3), espacialmente o químicamente aislados, compuestos de varios componentes celulares y que llevan funciones discretas. Estos módulos se pueden considerar elementos fundamentales en la organización celular, si están presentes en redes bioquímicas altamente integradas. Ravasz *et al.* [73] mostraron que redes metabólicas de 43 organismos distintos están organizadas en muchos *módulos* topológicos pequeños altamente conectados (ver sección 3.1.2, capitulo 3). Estos se combinan de manera *jerárquica* en unidades más grandes y menos cohesionadas, en las que su distribución de grado y coeficiente de agrupación (ver sección 2.2 capítulo siguiente) siguen una *ley de potencias* (figura 1.17). Dentro de *Escherichia coli*, la *modularidad jerárquica* descubierta se superpone estrechamente con funciones metabólicas conocidas. Mostraron que la arquitectura de red identificada puede ser genérica para la organización celular a nivel de sistema.

Esfuerzos como los descritos han propiciado un nuevo enfoque en la investigación biomédica post-genómica, en la que uno de sus objetivos clave es catalogar sistemáticamente todas las moléculas y sus interacciones dentro de una célula viva. Por lo que existe una clara necesidad de comprender cómo estas moléculas y sus interacciones determinan la función de esta maquinaria enormemente compleja, tanto de forma aislada

como acompañada por otras células. Así, los avances en biología indican que las redes celulares se rigen por leyes universales y ofrecen un nuevo marco conceptual que podría revolucionar nuestra visión de la biología y las patologías de las enfermedades en el siglo XXI. Con lo que contamos hoy en día con un enfoque teórico conocido como **Biología de Redes** [74].

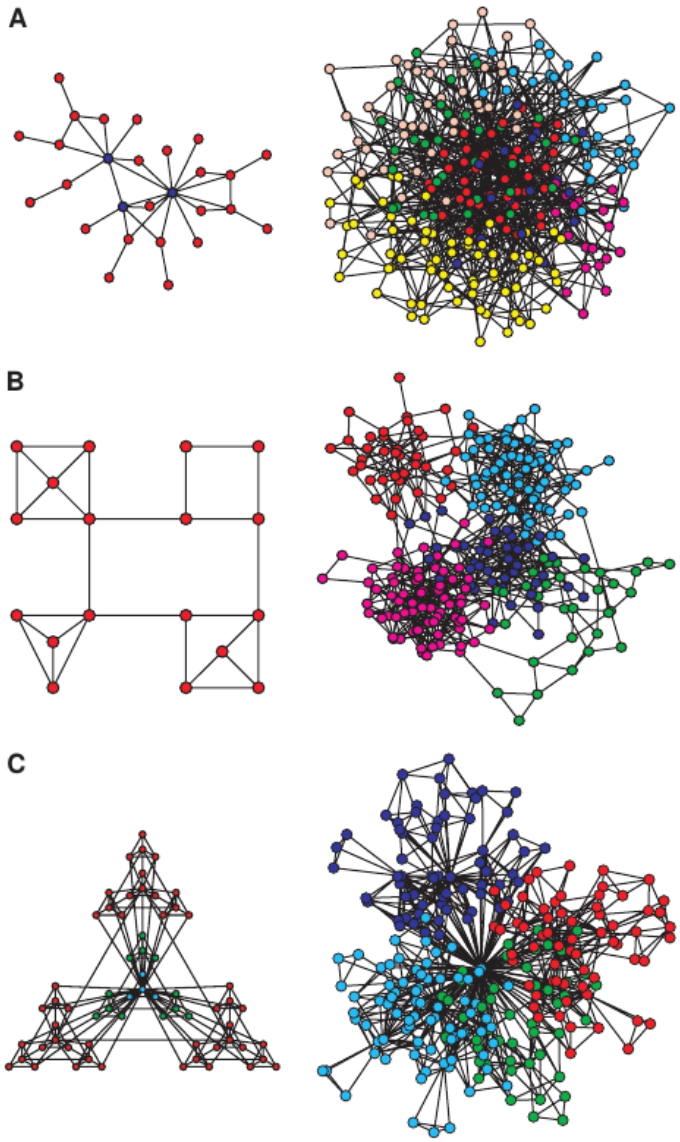


Figura 1.17: Una red compleja. A) Topología libre de escala, B) estructura modular y C) jerárquica

Sin embrago, el papel de las redes no termina aquí. En los últimos años, hemos aprendido que los efectos de red afectan cada vez más a todos los aspectos de la investigación biológica y médica, desde los mecanismos de la enfermedad hasta el descubrimiento de fármacos. Solo es cuestión de tiempo hasta que estos avances comiencen a afectar la práctica médica también, marcando el surgimiento de un nuevo campo que bien puede llamarse ***Medicina de Red***.

Redes aplicadas al estudio de enfermedades.

El creciente interés en la interconexión ha puesto de relieve una cuestión a menudo ignorada: las redes impregnan todo aspectos de la salud humana (figuras 1.18 y 1.24). Un ejemplo de esta tendencia son las redes sociales y su impacto en la propagación de la obesidad o los agentes patógenos, desde la gripe hasta el síndrome respiratorio agudo severo o el virus de inmunodeficiencia humana (VIH). El papel de las redes neuronales en diversas enfermedades psiquiátricas y neurodegenerativas es otro ejemplo.

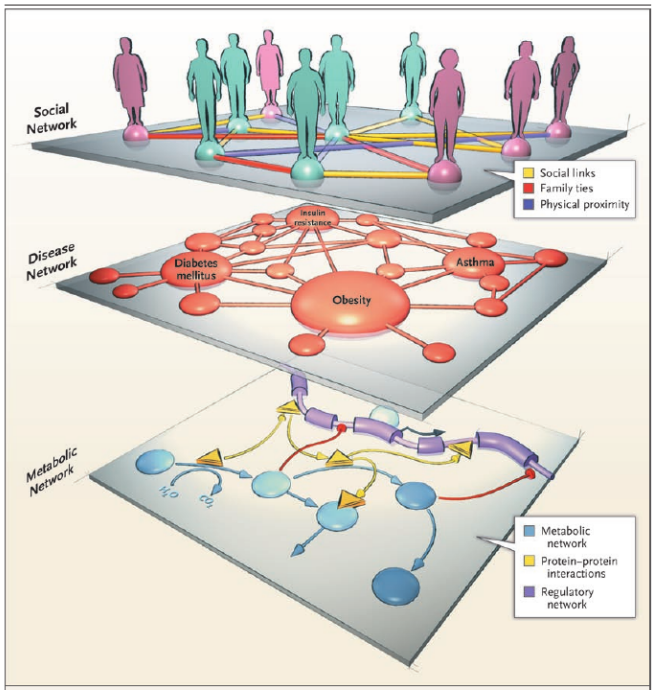


Figura 1.18: Redes complejas en diferentes escalas aplicadas a la biología y la salud.

La existencia de intrincados enlaces moleculares entre los componentes subcelulares y genes de una enfermedad plantea otra posibilidad: las enfermedades pueden no ser tan independientes entre sí como los médicos actuales las consideran. Así, para comprender los variados mecanismos de enfermedades, no es suficiente conocer la lista precisa de los

"genes de una enfermedad"; en su lugar, debemos tratar de trazar el diagrama detallado de los diversos componentes celulares que están influenciados por estos genes y productos genéticos. Tal pensamiento basado en la red puede proporcionar información sobre la patogénesis de varias enfermedades.

Por ejemplo, las enfermedades humanas, forman una red en la que dos enfermedades están conectadas si comparten al menos un gen (figura 1.19). En esta red de enfermedades, la obesidad tiene vínculos con otras enfermedades, como asma, lipodistrofia y glioblastoma. Por lo tanto, podríamos preguntarnos ¿podría un origen genético explicar el hecho de que la obesidad es un factor de riesgo para la diabetes? El concepto de red puede ayudarnos a responder la pregunta revelando conexiones entre enfermedades, que nos haga reconsiderar la forma en que las clasificamos y las separamos.

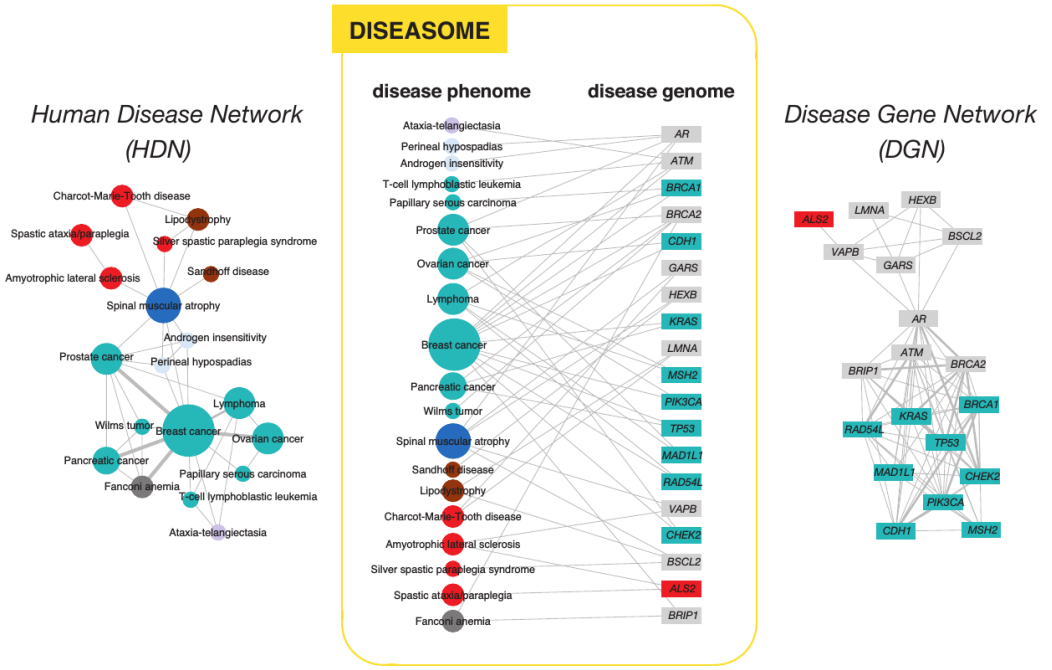


Figura 1.19: *Diseasome*: las enfermedades humanas, forman una red bipartita (sección 2.1.1.1, capitulo 2) en la que dos enfermedades están conectadas si comparten al menos un gen. Las proyecciones se muestran a la derecha (red de enfermedades) e izquierda (red de genes).

La red de enfermedades y genes unidas por asociaciones conocidas entre estos (figura 1.19), ofrece una plataforma para explorar en un único marco teórico de *redes complejas* todas las asociaciones conocidas de fenotipos y genes de enfermedades, lo cual puede

indicar el origen genético común de muchas enfermedades. Los genes asociados con trastornos similares muestran una mayor probabilidad de interacciones físicas entre sus productos y una mayor similitud de perfiles de expresión para sus transcripciones, lo que respalda la existencia de módulos funcionales específicos de enfermedades distintas (ver sección 3.1.2, capitulo 3). Goh *et al.* [75] encontraron que genes humanos esenciales son capaces de codificar proteínas concentradas y se expresan ampliamente en la mayoría de los tejidos. Esto sugiere que los genes de enfermedades también jugan un papel central en el interactoma humano (figura 1.20). En contraste, encontraron que la gran mayoría de los genes de enfermedades no son esenciales y no muestran tendencia a codificar proteínas concentradas, y su patrón de expresión indica que están localizados en la periferia funcional de la red. Asimismo, plantearon un modelo basado en selección que explica la diferencia observada entre genes esenciales y de enfermedad y también sugiere que las enfermedades causadas por mutaciones somáticas no deben ser periféricas, una predicción que confirmaron para genes de cáncer.

Así, dadas las interdependencias funcionales entre los componentes moleculares en una célula humana, una enfermedad rara vez es consecuencia de una anomalía en un solo gen, pero refleja las perturbaciones de la red compleja intracelular e intercelular que une los sistemas de órganos y tejidos (figura 1.18). Estos esfuerzos, se han concentrado en generar herramientas matemáticas para el estudio de éstas redes en lo que hoy conocemos como ***Medicina de Red*** [76]. Estas herramientas ofrecen una plataforma para explorar sistemáticamente no solo la complejidad molecular de una enfermedad particular, lo que lleva a la identificación de módulos y vías de enfermedad (ver sección 3.1.2, capitulo 3), sino también a las relaciones moleculares entre fenotipos patogénicos aparentemente distintos. Los avances en esta dirección son esenciales para identificar nuevos genes de enfermedades, para descubrir la importancia biológica de las mutaciones asociadas a enfermedades identificadas mediante estudios de asociación de genoma completo y secuenciación completa del genoma, y para identificar blancos farmacológicas y biomarcadores para enfermedades complejas.

A largo plazo, las redes pueden afectar todos los aspectos de la investigación médica y la práctica. De hecho, la cuestión fundamental de dónde se encuentra la función dentro de una célula está cambiando lentamente de un enfoque único en los genes a la comprensión de que detrás de cada función celular hay un módulo de red discernible que consiste en genes, factores de transcripción, ARN, enzimas y metabolitos (ver sección 3.1.2, capitulo 3). Esta comprensión nos hace ver a las enfermedades como el desglose de *módulos funcionales* seleccionados en lugar de como grupos únicos o pequeños de genes. Dados los muchos componentes de dichos *módulos funcionales* (ver sección 3.1.2, capitulo 3), existen diferentes caminos hacia la falla de los sistemas inductores de enfermedades; esto explica por qué a menudo muchos genes están relacionados con el mismo fenotipo de enfermedad. Del mismo modo, los efectos de los medicamentos no se limitan a las moléculas a las que se unen directamente; en cambio, estos efectos se pueden propagar a través de la red celular en la que actúan, causando efectos secundarios no

deseados. Por lo tanto, los efectos secundarios de los medicamentos son intrínsecamente fenómenos de red.

Asimismo, los sistemas biológicos complejos y las redes celulares pueden ser la base de la mayoría de las relaciones entre genotipos y fenotipos (figura 1.20). Por ejemplo Vidal *et al.* [77] explicaron por qué considerar las redes interactómicas son importantes para la biología; y asimsimo, cómo se pueden mapear e integrar entre sí, qué propiedades globales emergen de los modelos de redes interactivas y cómo estas propiedades pueden relacionarse con las enfermedades humanas (figura 1.16). De la misma manera, muchas enfermedades comunes, como el asma, la diabetes o la obesidad, implican interacciones alteradas entre miles de genes. Hoy en día, las técnicas de alto rendimiento (ómicas) permiten la identificación de tales genes y sus productos, pero la comprensión funcional es un desafío formidable. Los análisis basados en la red de datos ómicos han identificado módulos de genes asociados a enfermedades que se han utilizado para obtener un nivel de sistemas y una comprensión molecular de los mecanismos de la enfermedad (ver sección 3.1.2, capitulo 3). Por ejemplo, la alergia se usó como módulo para encontrar un nuevo gen candidato que fue validado por estudios funcionales y clínicos (figura 1.21). Tales análisis desempeñan un papel importante en la medicina de sistemas. Esta disciplina emergente que tiene como objetivo obtener una comprensión traslacional de los complejos mecanismos subyacentes a las enfermedades comunes. Gustafsson *et al.* [78] explicaron cómo los análisis basados en la red de datos ómicos, en combinación con estudios funcionales y clínicos, están ayudando a nuestra comprensión de enfermedades, además de ayudar a priorizar los marcadores diagnósticos o genes candidatos terapéuticos (figura 1.21).

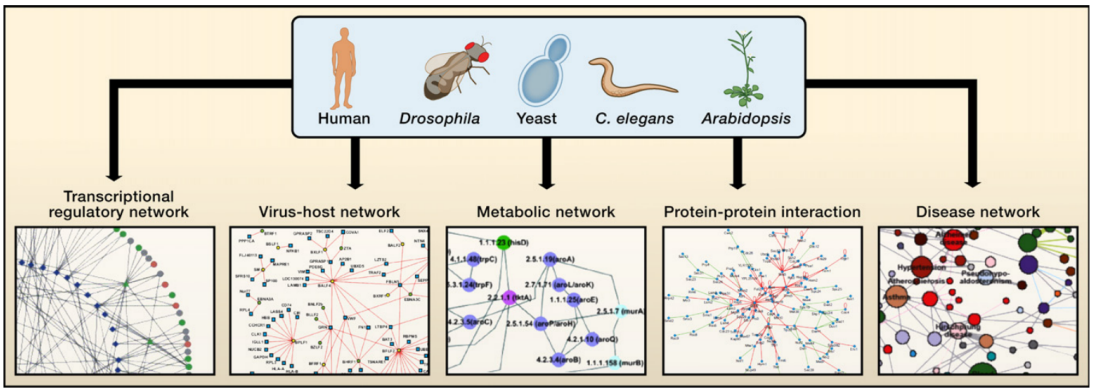


Figura 1.20: *Interactoma*: Diversas redes biologicas de diferentes tipos, en diversos organismos como levadura, gusano, mosca y planta, interactuan de manera conjunta.

El Proyecto del Genoma Humano ha revolucionado la búsqueda de genes, lo que ha provocado una explosión en el número de asociaciones detectadas entre los genes y los

fenotipos de la enfermedad. La belleza de los estudios de asociación genómica radica en su capacidad para cuantificar sus propias limitaciones. Por ejemplo, muchas de las mutaciones genéticas recién asociadas a la enfermedad representan solo una pequeña fracción de las ocurrencias de la enfermedad. Hay una tendencia a creer que el resto está oculto en más genes.

Así también, en la era post-genómica, la elucidación de la relación entre los orígenes moleculares de las enfermedades y sus fenotipos resultantes es una tarea crucial para la investigación médica. Zhou *et al.* [79] utilizaron una base de datos de literatura biomédica a gran escala para construir una red de enfermedades humanas basada en síntomas e investigamos la conexión entre las manifestaciones clínicas de las enfermedades y sus interacciones moleculares subyacentes. Encontraron que la similitud basada en síntomas de dos enfermedades se correlaciona fuertemente con el número de asociaciones genéticas compartidas y el grado en que interactúan sus proteínas asociadas. Además, la diversidad de las manifestaciones clínicas de una enfermedad puede estar relacionada con los patrones de conectividad de la red de interacción de proteínas subyacente. El mapa completo de las relaciones enfermedad-síntoma puede utilizarse como recurso para abordar cuestiones importantes en el campo de la medicina de sistemas, por ejemplo, la identificación de asociaciones inesperadas entre enfermedades, investigación de etiología de la enfermedad o diseño de fármacos.

Las enfermedades rara vez son el resultado de una anomalía en un solo gen, pero implican una cascada completa de interacciones entre varios procesos celulares. Para desentrañar estas complejas interacciones, es necesario estudiar las relaciones genotipo-fenotipo en el contexto de las redes de interacción proteína-proteína. Así, otro enfoque de la Medicina de Red es la observación de que las proteínas asociadas a enfermedades con el objetivo de dilucidar los mecanismos moleculares de la enfermedad humana. Tales enfoques se basan en la suposición de que las redes de interacción de proteínas se pueden ver como mapas en los que las enfermedades se pueden identificar con perturbaciones localizadas dentro de un vecindario determinado. La identificación de estos vecindarios, o módulos de enfermedades (ver sección 3.1.2, capitulo 3), es por lo tanto un requisito previo para una investigación detallada de un fenotipo patogéno particular. Ghiassian *et al.* [80] analizaron las propiedades de la red de de 70 enfermedades complejas y encontraron que las proteínas asociadas a enfermedades no residen dentro de comunidades localmente densas, en su lugar identificaron la importancia de la conectividad como la cantidad más predictiva. Esta cantidad inspiró el diseño de un algoritmo novedoso de Detección de Módulos de Enfermedad (DIAMOnD) para identificar módulos de enfermedades alrededor de un conjunto de proteínas de enfermedad conocidas (figura 1.22). Con este algoritmo validaron sistemáticamente las vecindades identificadas para un gran corpus de enfermedades.

Los genes compartidos representan una representación poderosa pero limitada de la relación mecanística entre dos enfermedades. Sin embargo, el análisis de las interaccio-

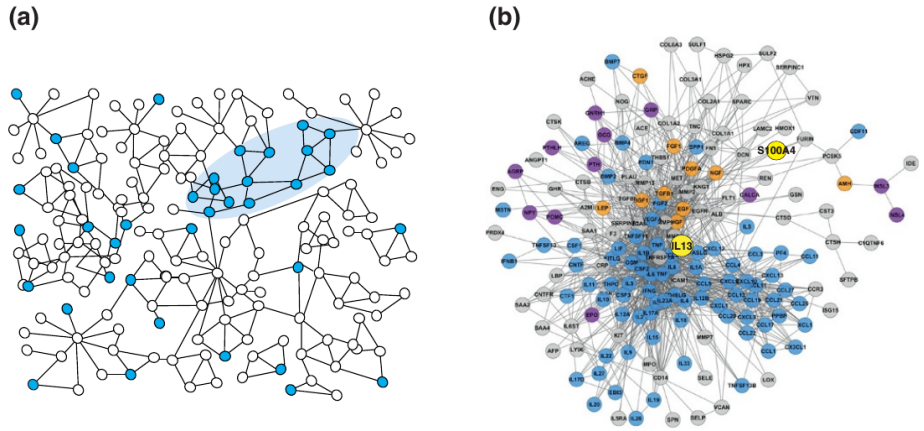


Figura 1.21: *Módulo de enfermedad.* (a) Los genes asociados a enfermedades (nódulos azules), tienden a co-localizarse en la red de interacción proteína-proteína humana (nódulos blancos), formando un módulo (óvalo azul). (b) Un módulo de enfermedad real de pacientes alérgicos.

nes proteína-proteína se ha visto obstaculizado por la incompletitud de los mapas de interactoma. Menche *et al.* [81] formularon las condiciones matemáticas necesarias para permitir que se observe un módulo de enfermedad (una región localizada de conexiones entre proteínas relacionadas con la enfermedad) (ver sección 3.1.2, capitulo 3). Solo las enfermedades con cobertura de datos que exceda un umbral específico tienen módulos de enfermedades identificables. La distancia basada en la red entre dos módulos de enfermedad reveló que los pares de enfermedades que se predice que tienen módulos superpuestos tenían similitud molecular estadísticamente significativa. Estas similitudes abarcan sus componentes proteicos, expresión génica, síntomas y morbilidad. También pueden identificarse enlaces a nivel molecular entre enfermedades que carecen de genes de enfermedades compartidas.

Sharma *et al.* [82] identificaron un módulo para asma, es decir, una vecindad local del interactoma cuya perturbación se asocia con el asma, y lo validaron por su relevancia funcional y fisiopatológica, utilizando enfoques tanto computacionales como experimentales (ver sección 3.1.2, capitulo 3). Encontraron que el módulo del asma está enriquecido con p-values de GWAS modestos en el contexto de variación aleatoria, y con genes expresados diferencialmente de fibroblastos asmáticos y normales tratados con un fármaco específico para el asma. El módulo de asma también contiene mecanismos de respuesta inmune que se comparten con otros módulos de enfermedades relacionadas con el sistema inmune. Además, utilizando diversos datos ómicos (genómica, expresión génica, respuesta a fármacos), identificamos la vía de señalización GAB1 como un nuevo modulador importante en el asma.

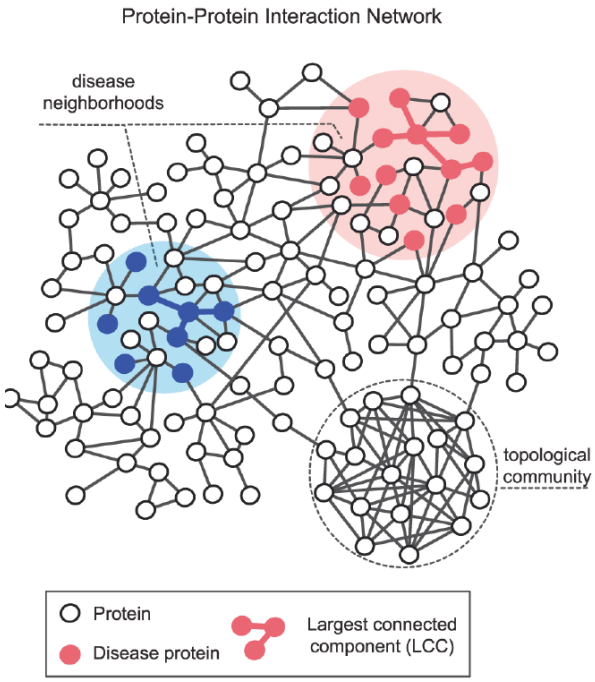


Figura 1.22: *DIseAse MOdule Detection (DIAMOnD)*. Proteínas asociadas con el mismo fenotipo tienden a localizarse en vecindarios específicos del Interactoma.

Históricamente, las enfermedades humanas se han diferenciado y categorizado según el sistema orgánico en el que se manifiestan principalmente. Recientemente, está surgiendo una visión alternativa que enfatiza que las diferentes enfermedades a menudo tienen mecanismos subyacentes comunes y patofenotipos intermedios compartidos. Dentro de este marco, la expresión de una enfermedad específica es una consecuencia de la interacción entre los endofenotipos relevantes y su entorno local basado en órganos. Ejemplos importantes de tales endofenotipos son la inflamación, la fibrosis y la trombosis y su papel esencial en muchas enfermedades en desarrollo. Ghiassian *et al.* [83] construyeron modelos de red endofenotipo y exploraron su relación con diferentes enfermedades en general y con enfermedades cardiovasculares en particular. Identificaron vecindarios locales (módulos) dentro del mapa interconectado de componentes moleculares, es decir, las subredes del interactoma humano que representan el inflamasoma, el trombosoma y el fibrosoma. Encontramos que estos vecindarios están superpuestos y están significativamente enriquecidos con genes asociados a enfermedades. En particular, también están enriquecidos con genes expresados diferencialmente relacionados con enfermedad cardiovascular. Asimismo, utilizando datos proteómicos, exploraron cómo la activación de los macrófagos contribuye a la comprensión de los procesos y respuestas inflamatorias. Los resultados del análisis muestran que las respuestas inflamatorias se inician dentro de la diafonía de los tres módulos endofenotípicos identificados.

Los genes que portan mutaciones asociadas con enfermedades genéticas están presentes en todas las células humanas; sin embargo, las manifestaciones clínicas de las enfermedades genéticas suelen ser altamente específicas de los tejidos. Aunque algunos genes de la enfermedad se expresan solo en tejidos seleccionados, los patrones de expresión de los genes de la enfermedad por sí solos no pueden explicar la especificidad tisular observada de las enfermedades humanas. Kitsak *et al.* [84] formularon la hipótesis de que para que una enfermedad se manifieste en un tejido en particular, se debe expresar una subred funcional completa de genes (*módulo de enfermedad*) en ese tejido (figura 1.23). Bajo esta hipótesis, llevaron a cabo un estudio sistemático de los patrones de expresión de genes de enfermedades dentro del interactoma humano. Encontraron que los genes expresados en un tejido específico tienden a localizarse en el mismo módulo del interactoma (ver sección 3.1.2, capitulo 3). Por el contrario, los genes expresados en diferentes tejidos están segregados en distintos vecindarios de red.

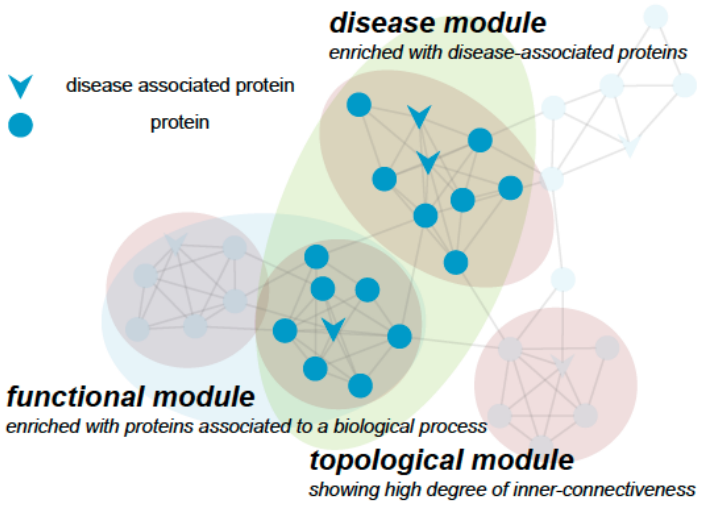


Figura 1.23: Módulos Funcionales: Procesos biológicos enriquecidos estadísticamente en módulos topológicos de una red de interacción de proteínas, pueden mapear a módulos en una red de enfermedades.

Más recientemente, Vinayagam *et al.* [85] caracterizaron la controlabilidad estructural de una gran red PPI humana dirigida que comprende 6,339 proteínas y 34,813 interacciones. Esta red permite clasificar las proteínas como "indispensables", "neutrales" o "dispensables", lo que se correlaciona con aumentar, no afectar o disminuir el número de nodos controladores en la red al eliminar esa proteína. Encontraron que el 21 % de las proteínas en la red PPI son indispensables. Curiosamente, estas proteínas indispensables son los objetivos principales de las mutaciones causantes de enfermeda-

des, los virus humanos y las drogas, lo que sugiere que la alteración de las propiedades de control de una red es fundamental para la transición entre estados sanos y estados de enfermedad. Esto sugiere que el análisis de controlabilidad es muy útil para identificar nuevos genes de enfermedades y posibles dianas farmacológicas.

Asimismo, Menche *et al.* [86] desarrollaron un marco para construir perfiles de perturbación personalizados para individuos, identificando el conjunto de genes que están significativamente perturbados en cada uno de ellos. Esto nos permite caracterizar la heterogeneidad de las manifestaciones moleculares de enfermedades complejas mediante la cuantificación de las similitudes y las diferencias del nivel de expresión entre pacientes con el mismo fenotipo. Se demuestra que a pesar de la alta heterogeneidad de los perfiles individuales de perturbación, los pacientes con asma, Parkinson y de Huntington enfermedad comparten un broadpool de los genes asociados de forma esporádica de la enfermedad, y que los individuos con solapamiento estadísticamente significativa con este grupo tienen la oportunidad de ser 80-100 % diagnosticado con la enfermedad. El marco desarrollado abre la posibilidad de aplicar datos de expresión génica en el contexto de la medicina de precisión, con implicaciones importantes para la identificación de biomarcadores, el desarrollo de fármacos, el diagnóstico y el tratamiento.

Naturalmente, el pensamiento basado en la red también puede explicar las influencias ambientales y sociales sobre la enfermedad. En este contexto, debemos comprender las interacciones humanas que abarcan los vínculos sociales y familiares, los contactos basados en la proximidad y las redes de transporte (figura 1.24).

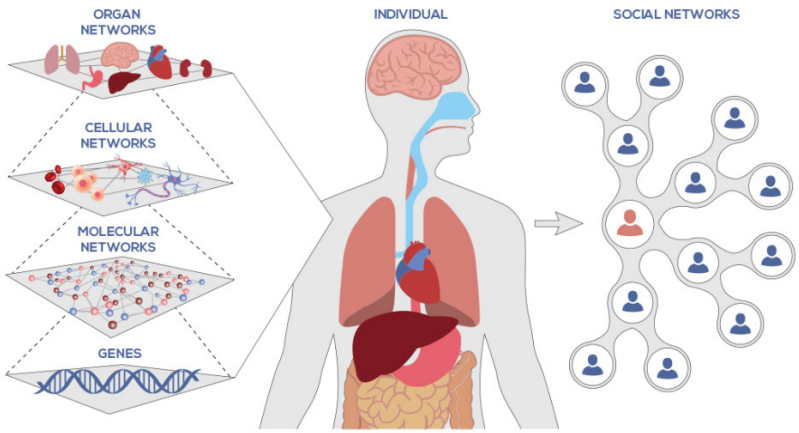


Figura 1.24: Integración bajo el enfoque de *Biología de Sistemas* de diferentes capas de redes biológicas (y sociales) en varias escalas. Un claro ejemplo de la aplicación de la *Ciencia de Datos* y la *Ciencia de Redes* a los sistemas biológicos

Aplicaciones al cancer.

El cáncer es una enfermedad altamente heterogénea y compleja, tal y como los muestran sus *sellos distintivos* o *Hallmarks* [87] (figura 1.25). Uno de los principales desafíos para su tratamiento es su la gran cantidad de variantes clínicas, fisiológicas y de supervivencia. Así ésta enfermedad es un candidato ideal para estudiarlo mediante el enfoque de *Biología de Sistemas* y de *Medicina de Redes*. Dado que hoy en día el cáncer se estudia mediante el uso de datos y herramientas específicos, como muestras de tumores de pacientes a partir de datos provenientes de tecnologías de alto rendimiento con especial atención a la caracterización del genoma del cáncer, así como íneas celulares de cáncer inmortalizadas, modelos de tumorigénesis en ratones, métodos de secuenciación de próxima generación, modelado computacional, etc.

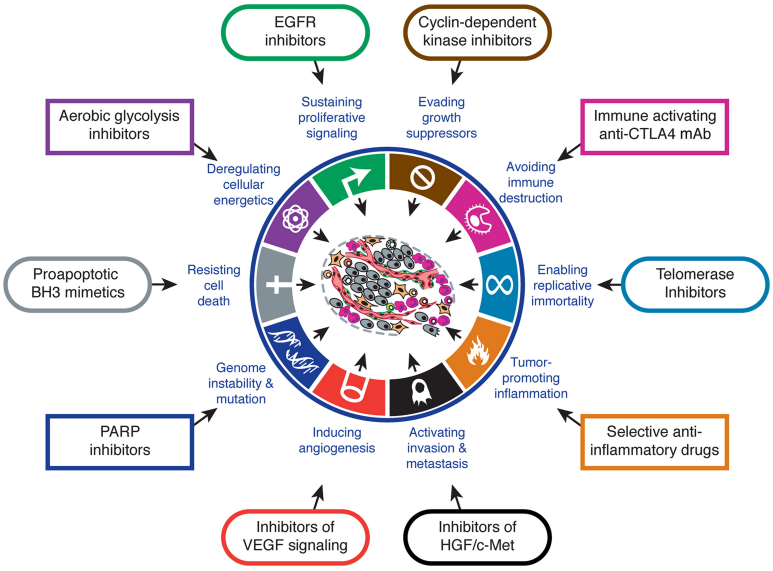


Figura 1.25: *Hallmarks of Cancer*: Los *sellos distintivos del cáncer* muestran lo compleja y heterogénea que es ésta enfermedad.

El objetivo a largo plazo de la biología de sistemas del cáncer es la capacidad de diagnosticarlo, clasificarlo y predecir mejor el resultado de un tratamiento sugerido, que es una base para la *medicina personalizada* contra el cáncer. De está manera los esfuerzos en la dirección de la *Medicina de Red* pueden lograr una mejor compresión de esta enfermedad. Así, elucidar las propiedades de las redes que modelan el cáncer, que distingan la enfermedad del estado celular saludable es, de vital importancia para obtener conocimientos a nivel de sistemas sobre los mecanismos de esta enfermedad, y en última instancia, para desarrollar mejores terapias.

Un ejemplo de o anterior es el de West *et al.* [88] quienes al integrar los datos de

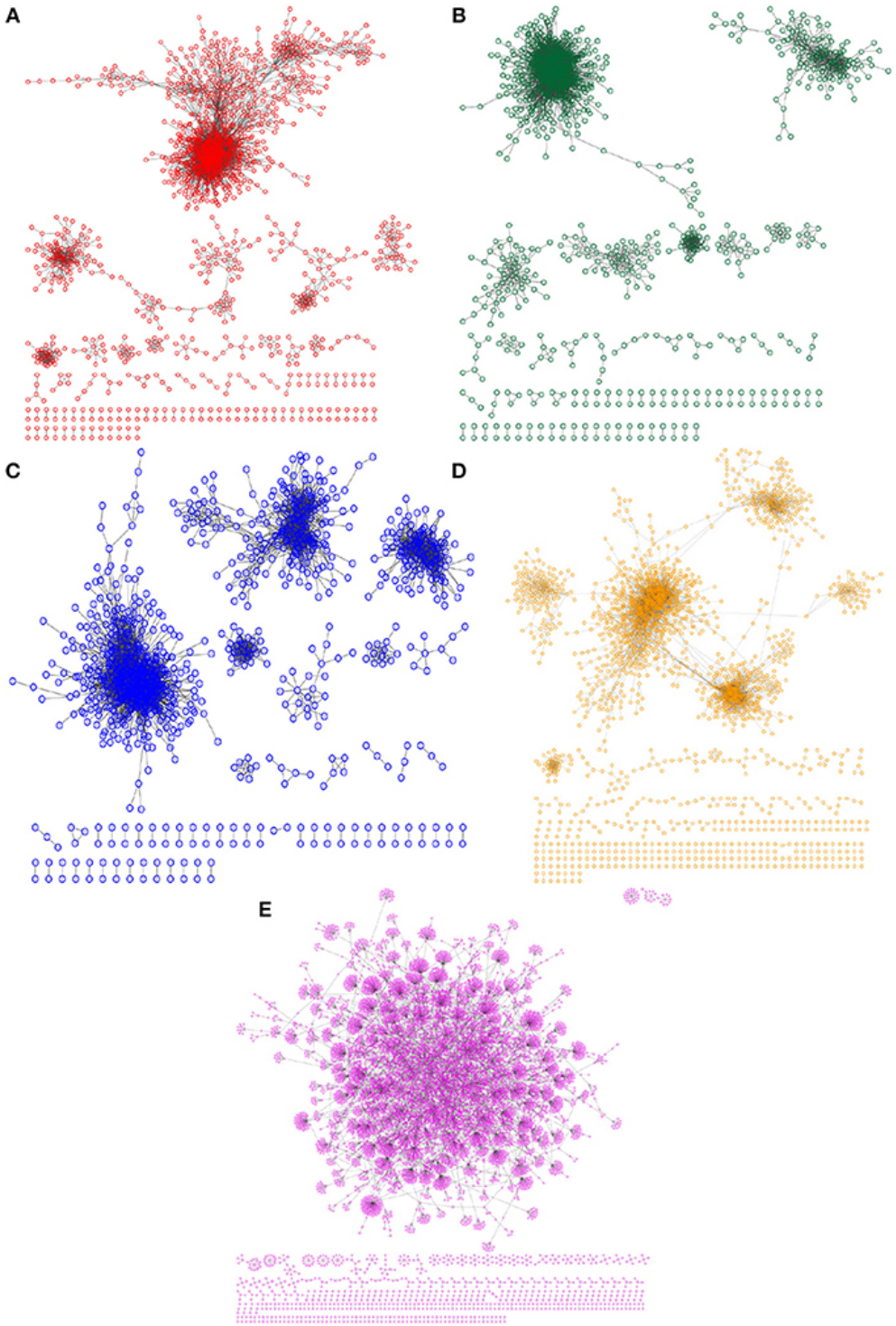


Figura 1.26: Redes asociadas inferidas a partir de biopsias de cáncer de mama. Se muestran las diferencias topológicas de las redes asociadas a los cuatro subtipos moleculares y la red de tejido sano.

expresión génica con una red de interacción de proteínas, demostraron que las células cancerosas se caracterizan por un aumento en la entropía de la red. Además, demostraron formalmente que las diferencias de expresión génica entre el tejido normal y el de cáncer están anticorrelacionadas con los cambios de entropía de la red local, proporcionando así un vínculo sistémico entre los cambios de expresión génica en los nodos y sus patrones de correlación local. En particular, encontraron que los genes que dirigen la proliferación celular en las células cancerosas y que a menudo codifican oncogenes están asociados con reducciones en la entropía de la red. Estos hallazgos pueden tener implicaciones potenciales para identificar nuevas dianas farmacológicas.

Redes de subtipos moleculares En otro ejemplo de redes complejas aplicadas a cáncer de mama De Anda *et al.* [69] construyeron redes transcripcionales para subtipos moleculares de cáncer de mama, a partir de perfiles de expresión génica (figura 1.26). Estás redes serán abordadas más adelante ya que parte de los resultados principales de este proyecto doctoral parten del estudio modular de éstas redes (ver capítulos 4 y 5).

Los autores mostraron que la heterogeneidad de los 4 subtipos de cáncer de mama se puede recuperar en la arquitecturas de redes transcripcionales.Asimismo dado que la estructura de la red está íntimamente ligada a la funcionalidad proponen que las redes inferidas son una representación de un programa de transcripción subyacente asociado a cada uno de los fenotipos estudiados.

Asimismo, encontraron una diferencia drástica entre las redes no tumorales y tumorales, del mismo modo que existe una diferencia drástica entre el tejido no tumoral y el tumoral. De hecho, encontraron una marcada diferencia en la estructura entre las redes de cáncer de mama y una red sin cáncer. La red transcripcional no tumoral está dominada por un componente gigante, mientras que las redes para cada uno de los subtipos de cáncer de mama estudiados presentan un mayor número de componentes disconexos. Lo anterior sugiere la existencia de una comunicación transcripcional generalizada en las células sanas, que se pierde y con una regulación fracturada y más autónoma en diferentes manifestaciones de cáncer.

Objetivo y planteamiento del proyecto de investigación.

De esta manera, una vez expuesto el marco principal en el que se desarrolla el trabajo doctoral, expondremos los objetivos principales del mismo. El proyecto pretende plantear una metodología computacional, matemáticamente robusta, para detectar módulos en redes de regulación genética de gran escala. Como vimos en este capítulo la modularidad es una constante en las redes biológicas desde las redes de regulación booleanas hasta las redes complejas de enfermedades; por lo tanto, poder detectar *módulos funcionales* de manera matemáticamente formal en estás redes es de vital importancia para el estudio de sistemas biológicos desde el enfoque de *Biología y Medicina de Red.*

Asimismo, las redes presentadas en este estudio (diferentes a las redes booleanas) pretenden ser un modelo robusto del *Programa de Regulación Genético*. Estás han sido inferidas bajo el enfoque de *Biología de Sistemas* a partir de datos experimentales de genoma completo disponibles en bases de datos públicas; en algunos casos generados mediante microarreglos de expresión. Así también, estás redes han sido analizadas bajo el enfoque de la *Biología de Red*, usando las propiedades topológicas de gran escala de las *Redes Complejas*, en particular su estructura modular, definidas formalmente en el sentido matemático.

Pregunta de Investigación y objetivos.

La pregunta principal que pretendemos responder mediante el proyecto de investigación es la siguiente: **¿los módulos encontrados mediante métodos computacionales en redes de regulación genética, corresponden con funciones biológicas?**

Asimismo, el objetivo principal del proyecto es **Identificar mediante métodos computacionales, módulos (comunidades) biológicamente funcionales en redes de regulación genética.**

Como objetivos particulares, nos hemos planteado:

- Proponer una metodología computacional y matemáticamente robusta para detectar módulos en redes de regulación genética de gran escala.
- Analizar redes de Regulación Genética inferidas a partir de datos experimentales de genoma completo disponibles en bases de datos públicas.
- Encontrar módulos y submódulos en redes de regulación génetica de genoma completo.
- caracterizar la estructura modular y jerárquica de redes de regulación genéticas de cáncer de mama.
- Modelar el programa regulatorio transcripcional de una manera formal matemática.

Panorama de la Tesis doctoral.

De está manera la tesis se divide en cuatro capítulos más. El siguiente está dedicado a detallar matemáticamente las propiedades topológicas de gran escala de las *Redes Complejas*, para posteriormente presentar ampliamente en el capítulo 3, los aspectos necesarios para encontrar módulos en redes complejas así como se revisarán varios de

los métodos principales al respecto. Así entonces en el capítulo 4 se expondrá nuestra metodología para encontrar *módulos funcionales* en redes de regulación genética.

Así entonces, los capítulos 2 y 3 son una descripción general de las redes complejas y de la modularidad en las mismas, el lector puede omitir dichos capítulos y entrar directamente al capítulos 4 donde exponemos nuestra propuesta metodológica y usar dichos capítulos como una referencia (a modo de libro de texto) para entender mejor este capítulo 4.

Finalmente en el capítulo 5 se expondrán y discutirán los resultados de aplicar esta metodología a diferentes casos de estudio. Dichos casos de estudio consisten en diferentes redes inferidas a partir de datos genómicos en particular de cáncer mama. Asimismo, en este capítulo se expondrán las conclusiones del proyecto de doctorado.

Capítulo 2

Redes Complejas.

En este capitulo abordaremos la *Teoría de Redes Complejas* de forma muy general y describiremos las características generales de una *red compleja*. Las ***redes complejas*** son conjuntos de muchos elementos conectados (*nodos*) que interactúan de alguna forma.

El capitulo es un compendio de los principales conceptos expuestos en revisiones extensas de autores ya clásicos y reconocidos en campo de la *Ciencia de Redes*, formados principalmente en *Física Estadistica* como Albert Lazlo Barabasi y Reka Albert [89], Mark E. J. Newman [90, 91], Sergey Dorogovtsev, y F.F. Mendes [92], Duncan Watts y Steven Strogatz [93], Stefano Boccaletti, Vito Latora y Yamir Moreno [94], Guido Caldarelli y Alessandro Vespignani [95]; por lo que se puede considerar una revisión de revisiones de la *Teoría de Redes*. Asimismo, un gran apoyo para la redacción de este capitulo fueron las notas sobre estos temas escritas por Maximino Aldana [96, 97] por lo que el capitulo en algunas partes se basa las mismas, las cuales son una lectura bastante recomendable para iniciarse en la *Ciencia de Redes*.

Las ***redes complejas*** tienen su fundamento matemático en la *teoría de gráficas* [98, 99] (*Graph Theory*), la cual a su vez tiene su origen en la famosa solución de Euler al problema de los puentes de Königsberg en 1735, convirtiéndose en uno de los pilares fundamentales de las matemáticas discretas. Durante el siglo XX la teoría de gráficas tuvo un desarrollo importante [100], destacándose la introducción de métodos probabilistas a principios de la década de 1960, especialmente los estudios de Erdös y Rényi [101, 102] de la probabilidad asintótica de conexión en una gráfica, que dio lugar a lo que se conoce como *teoría de gráficas aleatorias*, que ha sido una fuente fructífera de resultados teóricos. Otro aporte fundamental a la teoría de redes proviene de la Ciencias Sociales. Desde la década de 1930, los sociólogos señalaron la importancia de patrones de conexión entre las personas para la comprensión del funcionamiento de la sociedad. Los estudios típicos de redes sociales abordan cuestiones de centralidad y conectividad de individuos en la red.

Sin embargo a finales de la década de 1990 y principios del siglo XXI, hemos sido

testigos de un nuevo movimiento sustancial en la investigación de redes. La reciente posibilidad de reunir y analizar datos en una escala mucho mayor a la que antes era posible, a través de computadoras cada vez más potentes y bases de datos cada vez más grandes, ha hecho posible estudiar muchos sistemas complejos de interés científico, que se pueden representar como redes (conjuntos de nodos o vértices apareados por lineas o aristas). Los ejemplos incluyen redes de comunicación y distribución eléctrica, la Internet y la WWW (worldwide web), redes telefónicas, redes biológicas (metabólicas o tróficas), redes neuronales como la del nematodo *C. elegans*, y varias redes sociales como redes de coautoría y de citas de artículos académicos.

Hasta hace poco, los cálculos que involucran redes de millones de nodos no hubieran sido posibles sin la accesibilidad a estas bases de datos que delatan la topología de las redes del mundo real. Y que han transformado el estudio de gráficas de decenas o a lo más cientos de vértices, en el estudio de redes con miles o incluso millones de nodos. En este momento los físicos comenzaron a incorporar ideas de la *Física Estadística*, desplazando la atención del análisis de gráficas pequeñas y propiedades individuales de nodos o aristas, hacia la consideración de las propiedades estadísticas a gran escala (propiedades macroscópicas) de las redes. Dichos análisis estadísticos han puesto de manifiesto que la *estructura* de una enorme red de elementos dinámicos que interactúan entre si, siempre afecta la función y *dinámica*[103] de estos sistemas que se comportan de forma colectiva, en donde emergen propiedades globales a partir de las individuales y la arquitectura de la red, dando paso a lo que conocemos hoy como la teoría de redes complejas o bien solamente teoría de redes [89, 90, 92, 94, 95].

Así, la ***teoría de redes*** tiene como objetivos encontrar las propiedades estadísticas, tales como longitudes de camino (*path lengths*) y distribuciones de conexiones (*degree distributions*), que caractericen la estructura y el comportamiento de los sistemas en red, así como sugerir formas adecuadas de cuantificación estas propiedades. Y por otro lado crear modelos que puedan ayudar a entender el significado de dichas propiedades estadísticas y predecir cuál será el comportamiento de los sistemas en red al medir dichas propiedades estructurales [90]. De está manera los desarrollos teóricos sobre los efectos de la estructura y la dinámica de las redes sobre el comportamiento de sistemas han comenzado a rendir sus primeros frutos [94]. Unos de las más importantes propiedades estructurales es la modular o de comunidad en redes, la cuál abordaremos en el siguiente capítulo.

Definiciones generales con base en la teoría de gráficas.

Una red es un conjunto de muchos elementos conectados (*nodos*) que interactúan de alguna forma. A los nodos de una red (también llamados *vértices*) se les suele representar por los símbolos $v_1, v_2, \ldots, v_N$, donde N es el número total de nodos en la red. Si un nodo v_i está conectado con otro nodo v_j, está conexión se representa por una pareja ordenada

(v_i, v_j). Una gráfica se puede visualizar como un conjunto de puntos conectados por líneas, como se muestra en la figura 2.1. La definición formal de una *gráfica* (o bien si el número de nodos o vértices es muy grande, una red) es la siguiente:

Definición 1. *Una* ***gráfica*** G *es un par de conjuntos* V *y* E *(* $G = \{V, E\}$ *), donde,* $V = \{v_1, v_2, ..., v_N\}$ *es un conjunto de* ***vértices*** *y* E *es un subconjunto de* V^2 *(*$E \subseteq V \times V$*), tal que* $E = \{(v_i, v_j) := e_{ij} \mid v_i \wedge v_j \in V\}$*, es un conjunto de pares no ordenados de elementos de* V*. Los elementos de* E *se llaman* ***aristas*** *o enlaces y los dos vértices que identifican una arista se llaman puntos finales.*

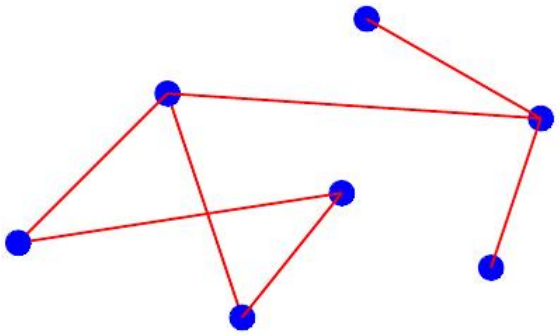


Figura 2.1: Un ejemplo sencillo de una gráfica (*binaria*) con siete vértices y siete aristas.

Por lo general, el número de vértices u *orden* de la gráfica se denota como $\boldsymbol{n}= |V|$ y el número de aristas o *tamaño* de la gráfica como $\boldsymbol{m}= |E|$. El tamaño máximo de una gráfica es igual al número total de pares de vértices no ordenados $m_{max} = \frac{n(n-1)}{2}$, cuando esto ocurre, es decir, que todos los vértices están conectados ente si, se dice que la gráfica es ***completa***.

Gráficas dirigidas y pesadas.

La *gráfica* G se llama ***no dirigida*** si para cada pareja $(v_i, v_j) \in E$ también existe la pareja $(v_j, v_i) \in E$. Sin embargo, las aristas pueden tener *dirección*, esto es, si existe una arista entre el vértice v_i y v_j $(e_{ij} \in E)$, pero no necesariamente existe una arista entre el vértice v_j y el v_i $(e_{ij} \notin E)$, es decir, si cada arista (v_i, v_j) es un *par* ***ordenado*** de vértices, entonces se tiene una ***gráfica dirigida*** (o digráfica); figura 2.2 panel A. Para este caso, e_{ij} es una arista dirigida de v_i a v_j, o bien, se dice que la arista comienza en v_i y termina en v_j.

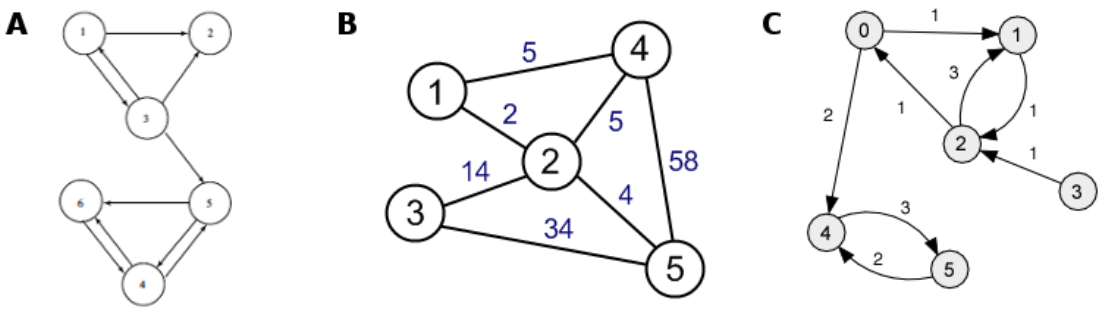


Figura 2.2: **Gráficas dirigidas y pesadas**. **A)** Una gráfica simple dirigida. **B)** Una gráfica pesada. **C)** Un gráfica dirigida y pesada.

Asimismo, se puede *ponderar* la interacción entre dos vértices asignando un número real o *peso* w_{ij} a la arista e_{ij} (por lo general $0 \leq w_{ij} \leq 1$), dando paso a lo que conocemos como gráficas (***pesadas***) 2.2 panel B. Muchas redes reales son *gráficas pesadas* y a diferencia de las gráficas dirigidas, algunas propiedades no suelen traducirse de forma inmediata del caso no pesado. El estudio de las gráficas pesadas es amplio y es recomendable revisar literatura especializada al respecto [70, 71, 104]. Por último, cuando una gráfica es no dirigida ni pesada se le conoce como una red binaria.

Subgráficas y otros tipos de gráficas relevantes.

Sub-gráficas. Una gráfica $G' = (V', E')$ es una sub-gráfica de $G = (V, E)$ si $V' \subseteq V$ y $E' \subseteq E$. Si G' contiene todas las aristas de G que unen vértices de V' se dice que la subgráfica G esta inducida o *abarcada* por V'. Una partición del conjunto de vértices V en dos subconjuntos S y $V - S$ se llama un ***corte***, el tamaño del mismo es el número de aristas de G que unen vértices de S con vértices de $V - S$.

Arboles. Por definición un árbol es una gráfica *acíclica*, es decir, que sólo puede haber un único camino desde un vértice a cualquier otro, si hubiera al menos dos caminos entre el mismo par de vértices, se formaría un ***ciclo***. El número de aristas de un árbol con n vértices es $n - 1$, si cualquiera de las aristas de un árbol se elimina, sería desconectado en dos partes; si una nueva arista se añade, habría al menos un ciclo. Así, los arboles son gráficas *mínimamente* conectadas y son muy importantes en la *teoría de gráficas* 2.3.

Multigráficas e hipergráficas. Formalmente las *gráficas* no incluyen *auto-conexiones*, es decir, aristas que conectan un vértice a sí mismo, ni aristas múltiples, *i.e.*, varias aristas uniendo el mismo par de vértices. Las gráficas con auto-conexiones y aristas múltiples

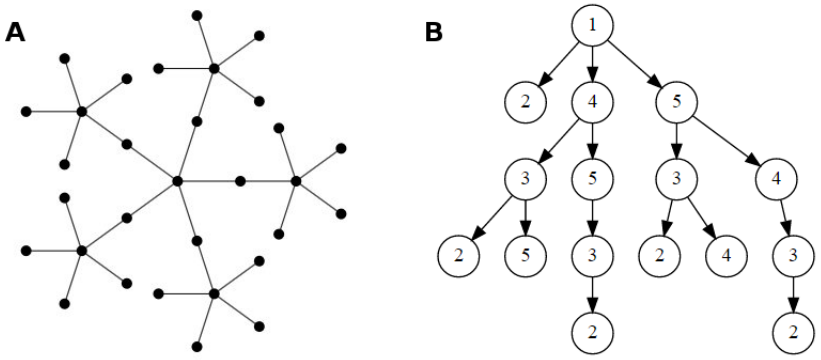


Figura 2.3: **Arboles**. **A)** Ejemplo de un árbol simple. **B)** Un árbol simple dirigido.

se llaman multigráficas. La generalización de gráficas que admiten aristas entre cualquier número de vértices (no necesariamente dos) se llaman ***hipergráficas*** [91].

Gráficas bipartitas y multipartitas. Una gráfica G es ***bipartita*** si el conjunto de vértices V se separa en dos subconjuntos **disjuntos** V_1 y V_2 llamados *clases* y cada arista une un vértice de V_1 con un vértice de V_2; pero no hay aristas que unan vértices dentro de la misma clase. La definición se puede ampliar a la de gráfica *r-partita*, donde hay r clases de vértice y ninguna arista une vértices dentro de la misma clase, en este caso se habla de gráficas ***multipartitas***.

Redes Complejas.

A partir de este momento y en el resto de este documento, entenderemos como una *red compleja*, una gráfica G con un orden de cientos, miles o decenas de miles e incluso millones de vértices y que presenta propiedades topológicas de gran escala. Asimismo por sencillez denotaremos al vértice v_i como el nodo i, de una red compleja. Aunque es importante señalar que tanto la notación de *gráfica* (o grafo) y *vértices* así como de *redes y nodos* se usa indistintamente en la literatura de redes complejas.

Representación y algunas propiedades matriciales de redes.

Matriz de Adyacencia.

Toda la información sobre la topología de una red de orden n está implícita en su ***matriz de adyacencia*** $\mathbb{A}$, que es una matriz de $n \times n$ cuyo elemento A_{ij} es igual a 1 si hay una arista que une los nodos i y j, o de lo contrario es cero. Debido a la ausencia de *auto-enlaces* de los elementos diagonales de la matriz de adyacencia son todos cero.

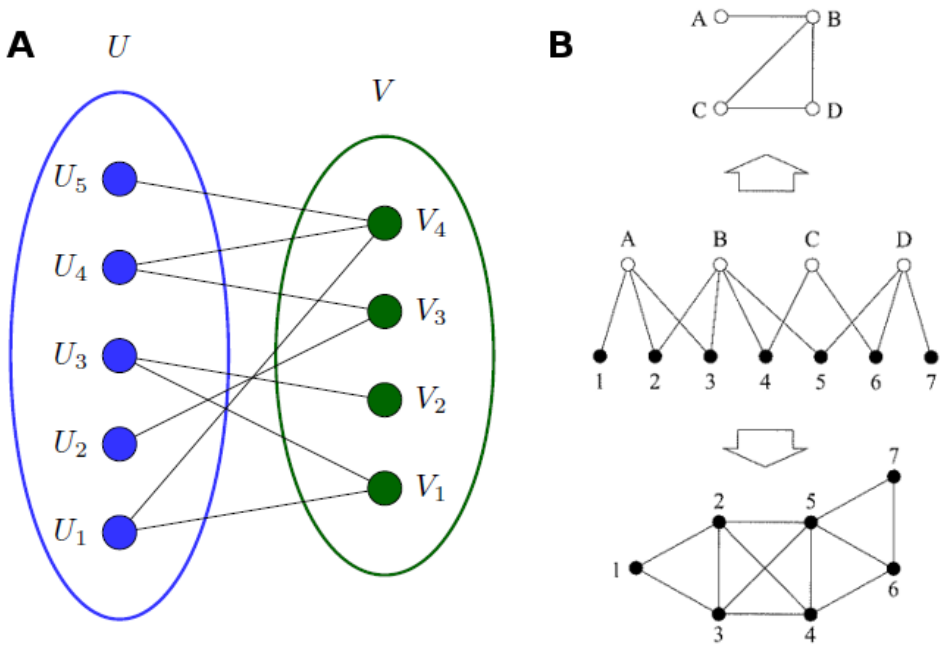


Figura 2.4: **Ejemplos de gráficas bipartitas**. **A)** Nótese la separación en dos subconjuntos *disjuntos* (clases). **B)** Las dos proyecciones únicas de una red bipartita. La parte central de esta figura muestra una red bipartita con cuatro vértices de un tipo (círculos abiertos etiquetados de A a D) y siete de otro (círculos rellenos, de 1 a 7). En la parte superior e inferior, se muestran las proyecciones de los dos conjuntos de vértices.

$$A_{ij} = \begin{cases} 1, & \text{si } (i,j) \in E, \\ 0, & \text{si } (i,j) \notin E. \end{cases} \tag{2.1}$$

Para redes no dirigidas $\mathbb{A}$ es una matriz *simétrica* y la suma de los elementos de la i-ésima fila o columna resulta en el número de conexiones (*grado*) del nodo i. Sin embargo para no dirigidas es *antisimétrica* $A_{ij} \neq A_{ji}$ y la suma de los elementos de la i-ésima fila resulta en el total de las conexiones que finalizan[1] en el nodo i y la suma de los elementos de la j-ésima columna es el total de conexiones que parten[2] desde el nodo j. Si las aristas se ponderan se define la matriz de ponderación (o matriz de pesos) $\mathbb{W}$, cuyo elemento w_{ij} expresa el peso de la arista entre los nodos i y j.

[1] *Grado de entrada*, siguiendo la convención de Mark E. J. Newman [91], véase la sección 2.2.1

[2] *Grado de salida*.

Propiedades espectrales.

El espectro de una red G es el conjunto de valores propios de su matriz de adyacencia $\mathbb{A}$. Las propiedades espectrales de las matrices asociadas a redes juegan un papel importante en el estudio de las mismas. Por ejemplo, para representar un proceso de difusión (o una caminata aleatoria) en una red es posible construir una matriz estocástica a partir de la matriz de adyacencia dividiendo los elementos de cada fila i por el grado de nodo i [1]. El espectro de dicha matriz permite evaluar el tiempo de difusión en la red, es decir, el tiempo que tarda en alcanzar la distribución estacionaria del proceso. Este último se obtiene calculando el *eigen*-vector de la matriz de transferencia correspondiente al *valor propio* más grande.

Matriz Laplaciana.

Otra matriz asociada a procesos de difusión en redes, es la Laplaciana. Esta se define para plantear y resolver la ecuación de un proceso difusión de algo (información por ejemplo) que fluye desde el nodo j hasta el nodo i a través de las aristas de la red. Formalmente, la matriz Laplaciana se define como: $\mathbb{L} = \mathbb{D} - \mathbb{A}$, donde $\mathbb{D}$ es la matriz diagonal cuyos elementos D_{ii} es igual al grado del nodo i y el resto de sus entradas es cero.

$$L_{ij} = \begin{cases} k_i, & \text{si } i = j, \\ -1, & \text{si } i \neq j \text{ y } (i,j) \in E \\ 0, & \text{en otro caso.} \end{cases} \tag{2.2}$$

La matriz Laplaciana $\mathbb{L}$, es una de las más estudiadas y encuentra aplicaciones en muchos contextos diferentes, como conectividad de redes [100], sincronización [105, 106], difusión [107] y el particionamiento de redes [108].

Una propiedad importante de la matriz Laplaciana es que por construcción, la suma de los elementos de cada fila es cero. Esto implica que $\mathbb{L}$ siempre tiene al menos un *valor propio* cero, correspondiente al *vector propio* cuyas componentes son todas uno: $\vec{v}_0 = (1, 1, \ldots, 1)$, dado que $\mathbb{L}\vec{v}_0 = \vec{0}$.

Así los vectores propios correspondientes a valores propios distintos son ortogonales entre sí y algo muy interesante es que $\mathbb{L}$ tiene tantos valores propios nulos como componentes conexos en la red. Por lo tanto, la matriz Laplaciana de una red conexa tiene un único valor propio cero y todos los otros son positivos. Los vectores propios (el espectro) de las matrices Laplacianas se utilizan regularmente para encontrar grupos y

[1] La transpuesta de esta matiz se conoce como *matriz de transferencia*

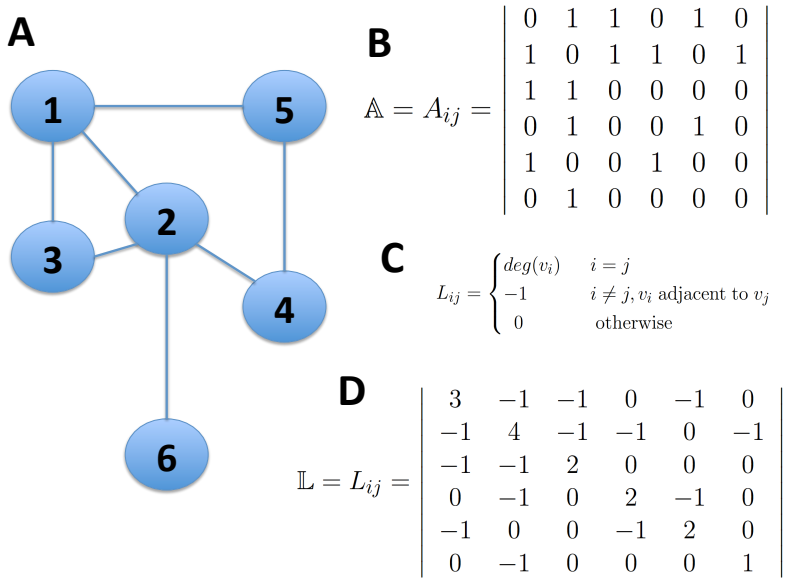


Figura 2.5: **Matriz laplaciana de una red**. **A)** Una pequeña red no dirigida **B)** Matriz de Adyacencia $\mathbb{A}$ que describe la conectividad de red de la red en el panel A; **C)** Definición de la matriz laplaciana de una red; **D)** matriz laplaciana $\mathbb{L}$ de la red en el panel A.

partitionar redes (véase capitulo siguiente (3.3.3).

En particular, el vector propio correspondiente al segundo valor propio más pequeño, llamado el vector de Fiedler [108, 109, 110], se utiliza para bipartitionar redes.

Centralidades básicas y caracterización de redes complejas.

Durante el estudio de la teoría de gráficas, el análisis visual era una excelente manera de ganar comprensión de su estructura, sin embargo, para una red de un millón o mil millones de nodos, este enfoque ya no es útil. El reciente desarrollo de métodos estadísticos para la cuantificación de las grandes redes (en general importados de la física estadística) [89, 111, 112] es en gran medida un intento de encontrar algo para jugar el papel desempeñado por el análisis visual de las gráficas del siglo XX. [1] Cuando

[1] Estos métodos responden a la pregunta: "How can I tell what this network looks like, when I can't actually look at it?" [90]

hablamos de miles o millones de nodos podemos encontrar propiedades estadísticas de las redes, las cuales han sido muy bien estudiadas. Dentro de las muy básicas podemos encontrar el *Clustering Coefficient*, la *Shortest path* entre cualesquiera par de nodos, pero lo que caracteriza la topología de una red es su distribución de grado, las cuales discutiremos muy brevemente a continuación.

Una vez definidos y mencionado los conceptos básicos de las redes y de la teoría de gráficas, en esta sección nos centraremos en revisar las propiedades que hacen a las redes un campo de estudio muy activo, nos referimos a las propiedades de gran escala de las *Redes Complejas*. En general el estudio de una red puede hacerse a partir del el análisis de su ***matriz de adyacencia*** (ver sección 2.1.2.1), sin embargo, se puede caracterizar una red a partir de un conjunto de características que son comunes a varios tipos de redes diferentes. En esta sección se exponen y discuten estas.

Conectividad y distribución de vecinos.

La característica más simple e inmediata de las redes es la conectividad de sus nodos. Dos nodos son *vecinos* (o adyacentes) si están conectados por una arista, y el conjunto de vecinos de un nodo i se llama vecindario. Así, el número $\boldsymbol{k_i}$ de vecinos del nodo i (es decir, el número de conexiones de i) se llama el ***grado*** (la conectividad) del nodo i. Y la *secuencia de grado* es la lista de los grados de los nodos de la red: $k_1, k_2, \ldots, k_n$. En términos de la matriz de adyacencia, para redes no pesadas $k_i = \sum_j A_{ij}$.

Para redes dirigidas se distinguen dos tipos de grado para un nodo i: el *grado de entrada* $\boldsymbol{k_i^{in}}$, que es el número de *aristas entrantes* (*i.e.* que terminan en el nodo i), así como el *grado de salida* $\boldsymbol{k_i^{out}}$ el número de sus *aristas salientes* (que parten desde i); de tal manera que $k_i = k_i^{in} + k_i^{out}$. En términos de la matriz de adyacencia, $k_i^{in} = \sum_j A_{ij}$ y $k_j^{out} = \sum_i A_{ij}$.

Asimismo, para redes pesadas, además del concepto de grado (k_i), existe el concepto de fuerza $\boldsymbol{s_i}$, que es la suma de los pesos de las aristas adyacentes al nodo i, es decir, $s_i = \sum_j w_{ij}$ [70, 71, 104].

Así, la conectividad de la red, es el promedio de los grados de todos los nodos y se denota como $\langle k \rangle$ donde:

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i \tag{2.3}$$

No todos los nodos de una red tienen el mismo grado y aunque quizá es la propiedad más simple de todas en la red, el estudio estadístico del conjunto de conexiones es muy importante. Así pues, la distribución de los grados de los nodos en una red, se puede

caracterizar por una función $\boldsymbol{P(k)}$.

Se puede definir $P(k)$ como la fracción de los nodos de la red que tienen grado k, o bien la probabilidad de que un nodo cualquiera seleccionado al azar tenga un grado k (*i.e.* exactamente k aristas), o bien k 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. Un diagrama de $P(k)$ para una red determinada, puede graficarse haciendo un histograma de los grados de nodos, este histograma es la distribución de grado para la red. La distribución de conexiones de los nodos ha sido ampliamente estudiada pues caracteriza la estructura de una red y delata diferencias que definen la topología la misma, se abundará más sobre esto en la sección 2.3.

Distancia mínima promedio entre nodos y *efecto de mundo pequeño.*

Otra propiedad importante que caracteriza a las redes complejas, es la longitud promedio del camino más corto (*geodésico*) entre dos nodos, la llamada *shortest path length.* El concepto de *camino* (*path*) se usa tanto para definir ***distancia*** como *conectividad* en redes:

Definición 2. *Un camino o* ***path*** *es una sub-red* $P \subset G(V, E)$ *de una red* G*, tal que* $P = \{V_P \subset V, E_P \subset E\}$*, con* $V_P = \{1, \ldots, l\}$ *y* E_P *es un conjunto ordenado de aristas tal que* $E_P = \{(1, 2), \ldots, (l-1, l)\}$*, donde* 1 *es el nodo inicial del path y para cada arista en* E_p *el punto final de cada arista es el nodo inicial de la siguiente, de tal manera que* l *es el nodo final de* P*, y* $\boldsymbol{l} = |E_P|$ *es la longitud del camino.*

Por ejemplo, un caso particular de camino es el *ciclo*, que es un *camino cerrado* donde los nodos inicial y final son el mismo y todas las aristas son distintas[1]. La potencia r-ésima de la matriz de adyacencia $\mathbb{A}^r$ el número de caminos que pasan por r aristas desde el nodo i al nodo j [91], así el numero total de ciclos de longitud 3 del nodo i, es decir el número de triángulos anclados al nodo i, se puede calcular con la entrada $\mathbb{A}^3_{ii}$ de la matriz de adyacencia.

Por otro lado, una red es *conexa* (está conectada) si, dado cualquier par de nodos, existe al menos un *camino* que va desde un nodo a el otro. Si no hay un camino entre dos nodos, la red está dividida en al menos dos sub-redes conectadas (figura 2.11). Cada subred máxima conectada de una red se llama *componente conexo.* Se profundizará sobre este concepto en la sección 2.2.5.2 .

[1] el ciclo más pequeño no trivial es un triángulo y su longitud es $l_c = 3$

Puede haber múltiples caminos con diferentes longitudes que conecten dos nodos, sin embargo de todos los caminos posibles siempre existe uno cuya longitud es la más corta de todas (*camino geodésico*). Así, la distancia entre dos nodos d_{ij} es el número de aristas a lo largo del *camino más corto* que conecta a los nodos i y j. En otras palabras, d_{ij} es el mínimo número de "saltos" que se tienen que dar desde un nodo i para llegar a otro nodo j en una red. En la red mostrada en la figura 2.6, existen varios caminos para llegar de i a j, sin embargo el camino más corto consiste de tres pasos (indicado con líneas gruesas).

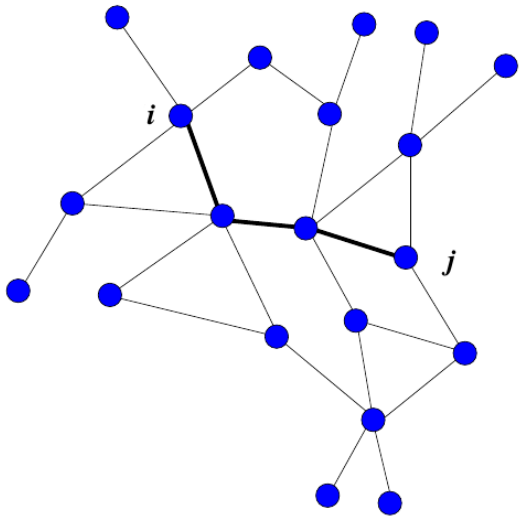


Figura 2.6: **Camino más corto en un componente conexo de una red**. Aunque existen varios caminos para llegar de i a j, el camino más corto está indicado con líneas gruesas. Figura tomada de las notas de Maximino Aldana..

De está manera, d_{ij} es la *longitud del camino más corto* o *distancia geodésica* entre i y j [89, 90, 92, 94, 95, 113].

Para un nodo dado i, la mayor distancia geodésica entre este nodo i y cualquier otro nodo en la red, se define como la **excentricidad del nodo ϵ_i**. El valor máximo de las excentricidades en toda la red $\boldsymbol{d = \text{máx}\{\epsilon_i \mid \forall\ i \in V\}}$, es llamado el **diámetro de la red** y el valor mínimo de todas de las excentricidades en toda la red $\boldsymbol{r = \text{mín}\{\epsilon_i \mid \forall\ i \in V\}}$, es llamado el **radio de la red**.

Si la red tiene más de un componente, convencionalmente la distancia entre nodos ubicados en componentes diferentes es infinita y por otro lado la distancia de un nodo a si mismo es cero.

Ahora bien, la longitud promedio de *caminos más cortos* en una red $\langle \boldsymbol{l} \rangle$ es el promedio de las longitudes mínimas d_{ij} sobre todos los pares de nodos (i, j) en la red. Así entonces, la longitud promedio de camino más corto (*shortest path length*) $\langle l \rangle$ de una red viene dada por[94, 113]:

$$\langle l \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij} \tag{2.4}$$

Otros autores [90] usan la siguiente definición:

$$\langle l \rangle = \frac{2}{n(n+1)} \sum_{i \geq j} d_{ij}$$

pero se puede mostrar que son equivalentes. Cabe señalar que calcular dicha *distancia mínima* d_{ij} entre cualesquiera par de nodos, no es trivial y existen muchos algoritmos para hacerlo [91], asimismo vale la pena subrayar que cuando la red es dirigida, d_{ij} no es necesariamente igual a d_{ji}. Asimismo, se puede notar que para redes con más de un componente la cantidad $\langle l \rangle$ diverge.

Para evitar dicha divergencia, se puede limitar la suma a sólo pares de nodos pertenecientes a los componentes más grandes. Otra alternativa usada en la literatura es definir $\langle l \rangle$ como la *media armónica* de las distancias geodésicas entre todos los pares, es decir, el recíproco de la media de los recíprocos, así los valores infinitos de d_{ij} no contribuyen en nada a la suma [90, 94]. Lo anterior conlleva a definir otras centralidades como la *eficiencia* (o comunicación) de la red (sección 2.2.5.1) y la *closeness* entre nodos que mencionaremos más adelante (sección 2.2.5.1). Asimismo, se puede obtener una medida de la relevancia de un nodo dado contando el número de geodésicas que lo atraviesan y definiendo la llamada *intermediación del nodo* (sección 2.2.5.1).

Efecto de mundo pequeño.

El hecho de que a pesar de su tamaño a menudo muy grande, en la mayoría de las redes haya un camino relativamente corto entre cualesquiera dos nodos, se conoce como *efecto de mundo pequeño* y ha sido estudiado y verificado directamente en un gran número de redes diferentes[89, 90]. Una de las primeras demostraciones directas de ésta característica (y tal vez, la más popular), es el famoso experimento del psicólogo social Stanley Milgram de 1967 [114], en el cual se le pidió a un grupo de personas pasar una carta a alguno de sus conocidos en un intento de alcanzar un individuo destino previamente designado. La mayoría de las cartas en el experimento se perdieron, pero alrededor de una cuarta parte llegó a la meta y pasó en promedio por las manos de sólo seis personas para hacerlo, dando paso al popular concepto de los "seis grados de

separación"[1].

El efecto de mundo pequeño, parece caracterizar la mayoría de las redes complejas y aunque es intrigante no es una indicación de un principio particular de organización, sin embargo tiene implicaciones obvias para la dinámica de procesos que tienen lugar en redes. Si se considera la difusión de información, o de cualquier otra cosa, a través de una red, el efecto de mundo pequeño implica que tal difusión será rápida, por ejemplo, si sólo toma seis pasos difundir un rumor de una persona a cualquier otra, entonces el rumor se extenderá mucho más rápido que si toma cien pasos, o un millón. Lo anterior tiene también implicaciones en el número de "saltos" que un paquete tiene que dar para ir de un computadora a otra a través de Internet o el tiempo que tarda una enfermedad en propagarse a través de una población [90].

En los últimos años el término "efecto de mundo pequeño", ha adquirido un significado más preciso: se dice que las redes muestran el efecto de mundo pequeño, si el valor de $\langle l \rangle$ escala logarítmicamente con el tamaño de la red para un grado promedio fijo, $\langle l \rangle \propto ln(N)$ [89, 90, 92, 94, 95].

Coeficiente de agrupamiento (*Clustering Coefficient*).

Otra cantidad local muy útil e importante en las redes complejas es el coeficiente de agrupamiento o agregación (*Clustering Coefficient*) [115], el cual se usa para describir las conexiones en el entorno más cercano a un nodo e indica el nivel de cohesión entre los vecinos del mismo.

Es común que en las redes sociales se formen pequeños grupos, como círculos de amigos o conocidos, en los que cada miembro conoce todos los demás miembros. Es muy probable que dos personas que tienen un colaborador mutuo también sean colaboradores entre si, o bien, en una red de amistades es muy probable que dos amigos de una persona también sean amigos uno del otro. Es decir si en una red el nodo A se conecta al nodo B y el nodo B con el nodo C, entonces hay una mayor probabilidad que el nodo A también se conecte al nodo C. Así, el *coeficiente de agrupamiento* de un nodo mide esta tendencia inherente a agruparse y tiene sus raíces en la sociología bajo el nombre de *transitividad* [89, 90, 92, 94, 95].

El *Clustering Coefficient* de un nodo C_i, caracteriza la "densidad" de conexiones en su entorno cercano. Mide que tanto los vecinos más cercanos de dicho nodo son vecinos cercanos el uno del otro, o bien, es la probabilidad media de que dos vecinos más próximos de un nodo sean vecinos entre si.

Es la relación entre el número de aristas que unen pares de vecinos del nodo i y el número total de aristas posibles, entre ellos. Según esta definición, C_i mide la probabi-

[1] aunque esa frase no figura en la escritura de Milgram

lidad de que un par de vecinos de i estén conectados (figura 2.7). Mas concretamente, C_i mide la fracción de tercias que tienen una tercer arista que completa un triángulo:

$$C_i = \frac{\text{triángulos conectados al nodo } i}{\text{todos los posibles triángulos conectados a } i}$$

Para una red no dirigida, fijémonos en un nodo i con k_i aristas que se conectan a otros k_i nodos. Si los vecinos del nodo i formaran una *subgráfica* completa (o *clique*) es decir que todos los vecinos están conectados entre si, entonces el número de todos los posibles triángulos conectados al nodo i pueden calcularse fácilmente como: $\binom{k_i}{2} = \frac{k_i(k_i-1)}{2}$, entonces C_i puede reescribirse como:

$$C_i = \frac{2E_i}{k_i(k_i-1)}$$

Donde E_i la fracción de todas esas posibles aristas que en realidad si están presenten en este subgrafo, es decir E_i son los triángulos centrados en i (que si existen). Para nodos con grado 0 o 1, en los cuales tanto el numerador como el denominador son iguales a cero, se tiene que $C_i = 0$.

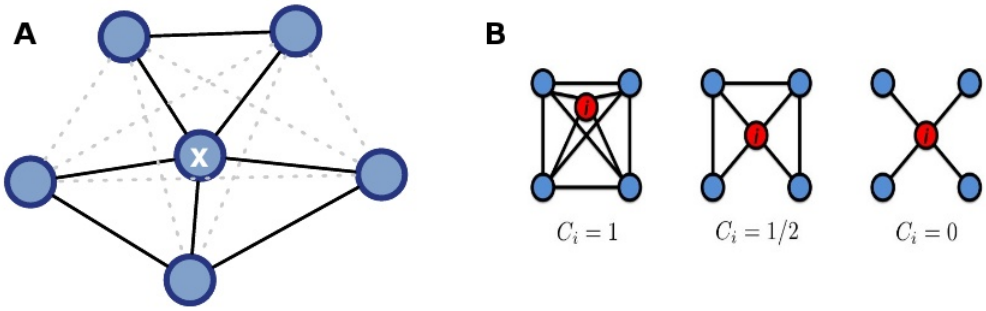


Figura 2.7: **Clustering Coefficient**. **A)** El *Clustering Coefficient* del nodo x es la relación entre los triángulos que si están formados con el nodo x y todos los posibles triángulos que se pueden formar con el nodo x; **B)** Diferentes valores de *Clustering Coefficient* para el nodo i.

De está manera $0 \leq C_i \leq 1$ y para *redes binarias* (no dirigidas ni pesadas) se puede calcular usando la potencia de la matriz de adyacencia (ver sección 2.2.2), por tanto:

$$C_i = \frac{\mathbb{A}^3_{ii}}{k_i(k_i-1)} = \frac{\sum_{m,l} a_{il}a_{lm}a_{mi}}{k_i(k_i-1)}$$

$$C_i = \frac{1}{k_i(k_i-1)}\sum_m\sum_l a_{il}a_{lm}a_{mi}$$

Promediando sobre todos los nodos se obtiene el ***coeficiente de agrupamiento*** de toda la red, $\langle C\rangle$ que viene dado por:

$$\langle C\rangle = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{2.5}$$

donde $0 \leq \langle C\rangle \leq 1$. Y puede usarse como una primera media de si la red exhibe una potencial estructura modular y/o jerárquica [73]. Es importante señalar que existen algunas definiciones alternativas de C_i [89, 90, 91, 92] y que para redes pesadas el calculo es diferente [70], por lo que se han propuesto generalizaciones [58, 116].

Distribución de Grado P(k).

Aunque el grado de un nodo es una cantidad local, la distribución de grado a menudo determina algunas importantes características globales de las redes. De está manera la distribución total de grados de los nodos de una red $P(k)$, es una de sus características estadísticas básicas [96, 97] . En la literatura existen muchos trabajos para modelar la distribución de vecinos $P(k)$ de redes reales ([89, 90, 91, 92, 93, 94, 95, 117]), sin embargo podemos mencionar algunos que han alcanzado gran relevancia por su sencillez y por ser pioneros en el área, pues varios modelos son modificaciones de estos, pero además los mismos determinan estructuras topológicas diferentes para las redes.

El más antiguo de estos es el propuesto por Erdös y Renyí [101, 102] el cual caracteriza a las denominadas ***redes (o gráficas) aleatorias*** (sección 2.3.1) que muestran una ***Topología de Poisson***:

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

Otro modelo muy importante, es el propuesto por Watts y Strogatz [115] (ver sección 2.3.2) que describe una ***red de mundo pequeño*** cuya topología varía a partir de un parámetro ***p*** del modelo entre una ***red ordenada*** (o lattice) y una red aleatoria con *Topología de Poisson*.

Los modelos anteriores describen la distribución de vecinos de redes con un número de nodos fijos, sin embargo también se han planteado modelos de crecimiento de redes, en el cual el número de nodos aumenta con el tiempo. Si el crecimiento de la red plantea que un nuevo nodo se enlaza de forma aleatoria como en el modelo de Erdös y Renyí, entonces la red resultante tendrá una ***Topología de Exponencial***:

$$P(k) = Ce^{-\alpha k}$$

Si por otro lado, el crecimiento plantea que un nuevo nodo se enlaza de forma aleatoria de manera preferencial a los nodos más conectados, se obtendrá una red con ***Topología Libre de Escala*** (ver sección 2.3.3), cuya distribución de vecinos toma la siguiente forma:

$$P(k) = Ck^{-\gamma}$$

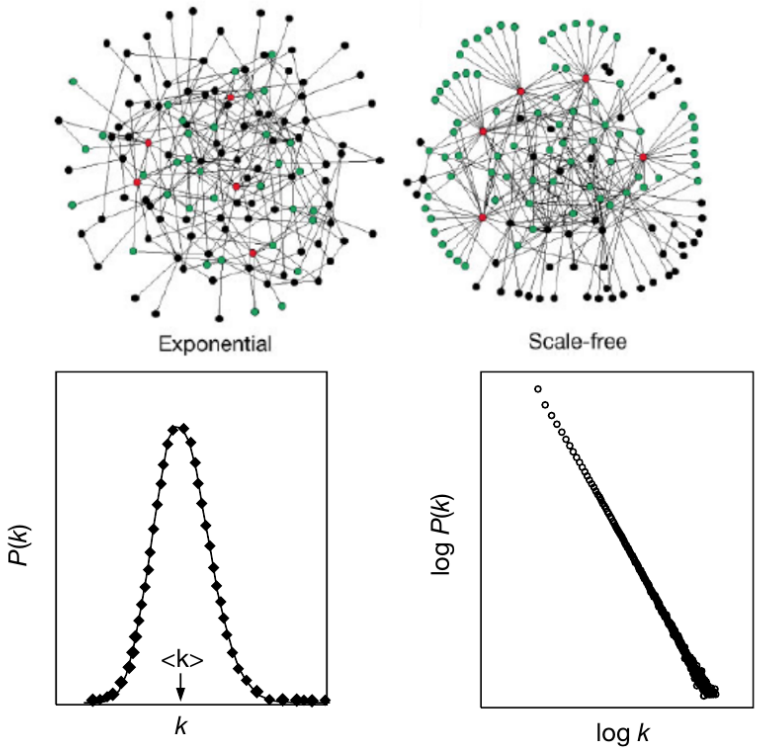


Figura 2.8: **Diferencias en la topología de redes**. A la izquierda se muestra una *red aleatoria* con topología exponencial (o *Topología de Poisson* y su distribución de grado (abajo). A la derecha se muestra una red con *Topología Libre de Escala* y (abajo) su distribución de grado graficada en $log(P(x))$ vs. $log(k)$.

Esta topolgía es muy importante, ya que como veremos más adelante, la mayoría de las redes reales que se encuentran en la naturaleza presentan este tipo de topología. Una discusión más detallada se estos modelos se puede ver más adelante en la sección 2.3.

Es muy importante mencionar que para el caso de redes pesadas también es posible estudiar la distribución de pesos, que puede diferir de la distribución de vecinos en su contraparate no pesada, por lo que extraer la información real de una red pesada no es un problema sencillo, pero que ha sido tratado por M. A. Serrano *et al.* [71].

Otras centralidades y propiedades de Redes Complejas.

Además de las propiedades de red presentadas, existen muchas otras medidas de centralidad que han sido propuestas y estudiadas en la literatura, así como otras que se derivan de las primeras. Así también existen propiedades generales de las redes entre las que se encuentran la *resistencia de red* o *los patrones de mezcla* entre otras. A continuación se exponen brevemente algunas de las más relevantes.

Centralidades de nodo.

Además del *grado* o el *Clustering Coefficient* de un nodo, existen otras centralidades destacadas en la literatura. Muchas de estás parten del concepto de distancia geodésica (*shortest path length*) entre nodos de la red y otras son generalizaciones del grado de un nodo y con el objetivo de medir mejor su "relevancia" dentro de la red.

Centralidad de intermediación (*betweenness centrality*). La *centralidad de intermediación* o betweenness centrality [90, 94], de un nodo i es el número de *caminos geodésicos* (véase sección 2.2.2) entre cualesquiera otros dos nodos, que pasan a través de i (figura 2.9). Más precisamente la *betweenness centrality* β_i de un nodo i se define como:

$$\beta_i = \sum_{j,k} \frac{\Lambda_{jk}(i)}{\Lambda_{jk}}$$

Donde Λ_{jk} es el número total de caminos geodésicos entre cualesquiera dos nodos j y k, mientras que $\Lambda_{jk}(i)$ es el número caminos geodésicos entre los nodos j y k que además pasan a través del nodo i.

Esta centralidad parece seguir una ley de potencia para muchas redes y también se puede ver como una medida de resistencia en la red (sección 2.2.5.2) que indica cuanto se incrementa la longitud promedio $\langle l \rangle$ cuando se elimina un nodo con un valor alto de *betweenness* [90].

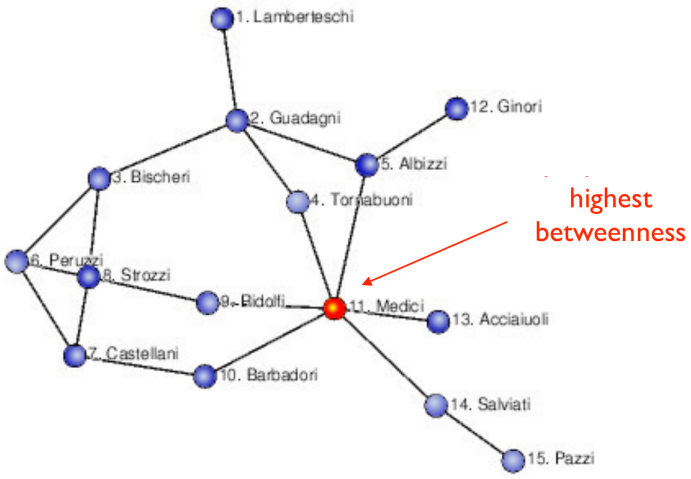


Figura 2.9: **Centralidad de intermediación (*betweenness centrality*)**. Se muestra un nodo con alta intermediación en una red personajes de la Florencia renacentista.

Cercanía (*Clossness*). Otra centralidad importante es la cercanía (o proximidad) γ_i, que mide qué tan lejos está un nodo de todos los demás en promedio.

$$\gamma_i = \frac{n-1}{\sum_i d_{ij}}$$

Esta cantidad toma valores bajos para los nodos que están separados de otros por una *distancia geodésica* corta en promedio. Estos nodos tendrán mejor acceso a la información de otros nodos o bien una influencia más directa sobre estos.

Eficiencia de la red (*Network efficiency*). La *distancia armónica* media entre un nodo y todos los demás $\langle \varepsilon_i \rangle$, se conoce como la *eficiencia* del nodo [118] y mide que tan eficientemente puede fluir información de cualquier tipo entre cualquier par de nodos. Así, la eficiencia entre cualquier par de nodos se define como: $\varepsilon_{ij} = \frac{1}{d_{ij}}$, donde d_{ij} es la *longitud del camino más corto* entre i y j. Si la distancia entre ambos nodos es grande la *eficiencia* es poca y viceversa. De ésta manera, la *eficiencia* promedio del nodo i viene dada por:

$$\langle \varepsilon_i \rangle = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

y por tanto la *Eficiencia* de la red es:

$$E = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}$$

Así como la centralidad de intermediación, la *Eficiencia de la red* también puede verse como una medida de la resistencia de la red (sección 2.2.5.2) que indica cuánto efecto tendrá la eliminación de un nodo sobre la transmisión de información en la red.

Centralidad de vector propio (*Eigenvector centrality*). La *centralidad de eigenvector* es una medida más refinada de la influencia de un nodo en una red y es una extensión natural de la centralidad del grado:

$$k_i = \sum_j A_{ij}$$

Podemos pensar en la centralidad de grado como una medida de "importancia" de un nodo. Por lo tanto, cualquier nodo en la red recibe una puntuación relativa por cada vecino que tenga en la red. Sin embargo, no todos los vecinos son equivalentes. En muchas circunstancias, la importancia de un nodo en una red se puede incrementar al tener conexiones con otros nodos que son también son "importantes" (centrales). Este es el concepto detrás de la *centralidad de vector propio*.

La centralidad de vector propio $\vec{x}$, se obtiene al calcular el *valor propio* más grande λ_1 de espectro de la matriz de adyacencia $\mathbb{A}$ de la red:

$$\mathbb{A}\vec{x} = \lambda_1\vec{x}$$

En otras palabras, el vector limitante de centralidades $\vec{x}$ es simplemente proporcional al vector propio líder de la matriz de adyacencia. Así la *eigencentralidad* x_i del nodo i es proporcional a la suma de las centralidades de los vecinos de i limitada por el *valor propio* más grande λ_1:

$$x_i = \frac{1}{\lambda_1}\sum_j A_{ij}x_j$$

Así, en lugar de atribuir a los nodos sólo un punto por cada vecino, la *centralidad de eigenvector* le otorga a cada nodo una puntuación proporcional a la suma de las puntuaciones de sus vecinos. Así las conexiones con nodos de alta puntuación contribuyen más a la puntuación del nodo en cuestión que las conexiones de los nodos de baja puntuación. La *centralidad de Katz* y el *PageRank de Google* son variantes de la centralidad de vector propio que se plantean como correcciones de la misma para redes dirigidas.

Centralidad de Katz. Para redes dirigidas, un nodo puede tener una centralidad de vector propio igual a cero si sólo tiene aristas salientes o bien si el vecino de su única arista entrante también tiene una centralidad de vector propio igual a cero. Así entonces se puede corregir agregando una cantidad arbitraria β de la siguiente manera:

$$x_i = \alpha\sum_j A_{ij}x_j + \beta$$

Donde α y β son constantes positivas que balancean el valor de la centralidad x_i. El primer termino de la ecuación es claramente la centralidad de vector propio y el segundo termino β "agrega centralidad" al nodo corrigiendo el problema. A esta modificación se le conoce como *centralidad de Katz* [119] y fue introducida en 1953 dentro del estudio de redes de citas.

PageRank. Ahora bien, es posible definir una variación de la centralidad de Katz en la que el valor que que un nodo i obtiene de sus vecinos sea proporcional a su centralidad dividida por su *grado de salida*. Entonces los nodos que apuntan a muchos otros aportan sólo una pequeña cantidad de centralidad, incluso si su propia centralidad es alta.

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

Esta centralidad se conoce comúnmente como PageRank [120], que es el nombre comercial dado por Google y que utiliza como parte central de su tecnología para clasificar los sitios en sus búsquedas. Esta centralidad puede reescribirse como:

$$\rho_i = \delta \sum_j \frac{\rho_j}{k_j^{out}} + \frac{1-\delta}{n} \tag{2.6}$$

Donde δ es el llamado *factor de amortiguamiento* (damping factor), para el cual Google usa un valor de $\delta \approx 0.85$. Y se puede notar que cuando la red es no dirigida, el *PageRank* de cada nodo es proporcional a su grado.

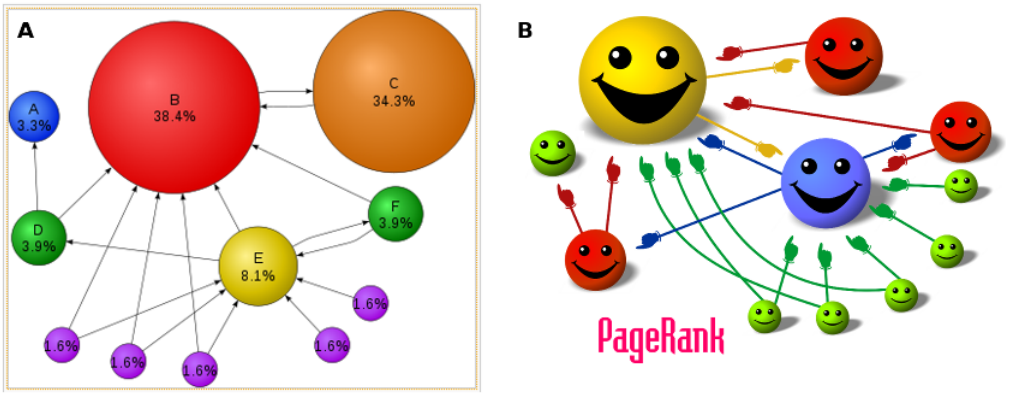


Figura 2.10: **Esquemas representativos de la centralidad de PageRank**. El tamaño de los nodos es proporcional a su *PageRank* y este a su vez toma en cuenta las conexiones entrantes de ponderando las de mayor *PageRank*. Nótese como el nodo C del panel A, tiene un PageRank alto a pesar de solo tener conexión con el nodo B.

Propiedades de Redes.

Además de las propiedades de red presentadas, existen muchas otras medidas de centralidad que han sido propuestas y estudiadas en la literatura, así como otras que se derivan de las primeras. Así también existen propiedades generales de las redes entre las que se encuentran la *resistencia de red* o *los patrones de mezcla* entre otras. A continuación se exponen brevemente algunas de las más relevantes.

Resistencia de red. Una propiedad que ha sido objeto de atención en la literatura, es la resistencia de las redes a la eliminación de sus nodos.

Dado que la mayoría de las redes basan su función en su conectividad y por lo tanto en la existencia de caminos que hay entre pares de nodos. Si se eliminan nodos de una red, la longitud típica de estos caminos se incrementan, y en última instancia, pares de nodos se desconectan y la comunicación entre ellos a través de la red sería imposible. Hay diferentes formas de eliminar nodos y redes diferentes muestran varios grados de resistencia a tal remoción de nodos. Por ejemplo, se podrían eliminar nodos aleatoriamente de una red, o se podría apuntar a alguna clase específica de nodos, tales como aquellos con los grados más altos.

La *resistencia de la red* es de particular importancia en epidemiología, en la que "la eliminación" de nodos en una red de contagios puede corresponder, por ejemplo, a la vacunación de individuos en contra de la enfermedad en cuestión. Dicha vacunación no sólo impide que los individuos vacunados contraigan la enfermedad, sino también puede destruir rutas entre personas por las cuales la enfermedad puede difundirse. De ésta manera, al estudiar la *resistencia de la red* de contagios, se puede tener un efecto de amplio alcance y una mejor eficiencia al diseñar estrategias de vacunación.

De igual manera la resistencia de la red puede tener aplicaciones en redes tecnológicas, al identificar posibles blancos que puedan provocar una caída grave en la comunicación de la red o bien sean susceptibles de ataques.

Componentes conexos y la *Isla* Gigante. Otro concepto importante muy abordado en la literatura, es el de los componentes (o islas) de una red.

Como vimos anteriormente (sección 2.2.2), una red puede contener partes desconectadas. Según la definición 1, una red es un conjunto de nodos entre los que existen algunas conexiones. Sin embargo no implica que todos los nodos deban estar conectados entre si, o que todos los nodos deban tener conexiones. En la red pueden existir nodos aislados, así como grupos de nodos que estén conectados entre sí pero que no estén conectados con el resto de la red, a estos conjuntos de nodos se les llama *componentes* o *islas*. La figura 2.11 muestra un ejemplo de una red compuesta de varias islas.

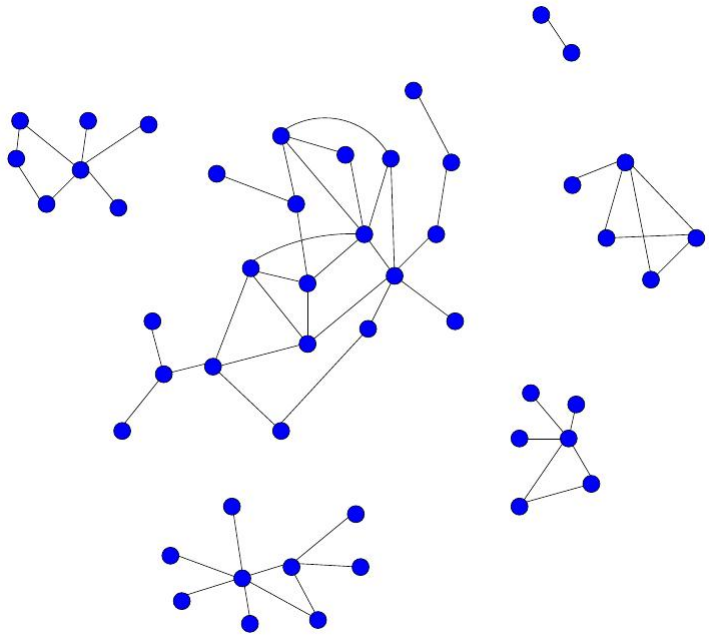


Figura 2.11: **Varios *componentes* de una red**. Si no existe un camino entre algún par de nodos, la red se divide en *componentes*, a veces llamados islas. Figura tomada de las notas de Maximino Aldana..

Cuando se plantea un modelo de crecimiento (secciones 2.3.1.2 y 2.3.3) en la que una red crece a un "tamaño infinito" (*límite de red grande*[1]); si el tamaño relativo del componente (o isla) conexa más grande respecto al resto de la red se aproxima a un valor distinto de cero, *i.e.* el componente más grande es casi toda la red, se dice que el sistema está por encima del *umbral de percolación* (ver sección 2.3.1.1), y a este grupo se le denomina el *componente conexo (o isla)* ***gigante*** de la red que puede denotarse como S_∞. En este caso, el tamaño de los demás componentes más próximos en tamaño, son muy pequeñas en comparación con el componente gigante para una red suficientemente grande.

Las islas en una red pueden tener diferentes tamaños, que van desde 1 (un sólo nodo que no está conectado con nadie) hasta el tamaño de toda la red (todos los nodos están conectados con algún otro en la red), en cuyo caso la red consiste de una sola isla, que es ella misma.

Es importante enfatizar que el hecho de que una isla no esté conectada al cuerpo

[1]también llamado límite termodinámico

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.

En algunas redes el tamaño del componente más grande es una cantidad importante. Por ejemplo, en una red de comunicación como Internet el tamaño del componente más grande representa la fracción más grande de la red dentro de la cual es posible la comunicación y por lo tanto es una medida de la efectividad con la que la red realiza su trabajo. También se suele medir el tamaño del segundo componente más grande de la red. En redes muy por encima de la densidad en la cual se forma un componente gigante por primera vez, se espera que el componente gigante sea mucho mayor que el segundo más grande.

Patrones de mezcla (*Mixing Patterns*). Profundizando un poco más en la estructura de red, nos podemos preguntar ¿qué nodos se emparejan con cuales otros? En algunas redes hay tipos diferentes de nodos, y la probabilidad de conexión a menudo depende de esos tipos. Por ejemplo, en una red trófica que representa qué especies se comen a cuales en un ecosistema, se pueden ver nodos que representan plantas, herbívoros y carnívoros. Muchas aristas vinculan plantas con herbívoros, y otras más a herbívoros con carnívoros. Sin embargo, hay pocas aristas que unen herbívoros con otros herbívoros o carnívoros con plantas. Lo mismo se puede ver en una red de regulación genética, con genes y *miRNAs*.

En las redes sociales este tipo de vinculación selectiva se llama "mezcla selectiva" (*assortative mixing*) u homofilia, donde ejemplo clásico es la *mezcla selectiva* por etnia. Esta propiedad también ha sido ampliamente estudiada en epidemiología. Y en la literatura sobre ecología se observa el término "correspondencia selectiva" (*assortative matching*) para referirse al mismo fenómeno, sobre todo en referencia a la elección de pareja entre animales.

La *mezcla selectiva* puede ser cuantificada por un "coeficiente de selectividad", que se puede definir en un par de de maneras diferentes. Esta cantidad es 0 en una red mezclada al azar y 1 en una red con una mezcla selectiva perfecta [90, 121].

Correlaciones de grado. Un caso particular de *mezcla selectiva*, es la mezcla conforme al grado, que es una propiedad escalar de los nodos. Esta propiedad es conocida comúnmente como *correlación grado*, y mide si nodos de alto grado se conectan preferente con otros nodos de grado alto, o bien prefieren adherirse a los de grado bajo. Resulta ser que ambas situaciones se observan en algunas redes.

El caso de la mezcla selectiva por grado es de particular interés ya que, dado que el grado en sí es una característica topológica de la red; dichas correlaciones pueden dar lugar a algunos efectos interesantes de la estructura de la misma. Al igual que el *coeficiente de selectividad*, en la literatura se han propuesto varias formas de cuantificar las correlaciones de grado. Por ejemplo, puede calcularse mediante el coeficiente de correlación de Pearson de los grados en ambos extremos de una arista; con lo que se obtiene solo número que debería ser positivo para redes mezcladas selectivamente y negativo en las no selectivas [90, 121].

Navegación en una red. Un problema ampliamente abordado en la literatura de redes complejas, es el de navegabilidad en redes, el cual trata sobre como encontrar vías rápidas para transmitir información entre dos puntos lejanos en la red. Si fuera posible construir redes artificiales que fueran fáciles de navegar de la misma manera que las redes sociales parecen serlo, sería posible usarlas para construir estructuras de bases de datos eficientes o mejores redes de computadoras.

El célebre experimento de Stanley Milgram [114] (ver sección 2.3.2), mostró que existen caminos cortos a través de las redes sociales entre individuos aparentemente distantes. Sin embargo, también muestra que la gente es buena encontrándolos, lo cual, tal vez es un resultado aún más sorprendente que la existencia de los caminos. Los participantes en el estudio de Milgram no tenían ningún conocimiento especial de la red que conectaba a la persona objetivo. En general, la mayoría de las personas sólo conocen quienes son sus amigos y tal vez algunos de los amigos de sus amigos. No obstante fue posible entregar un mensaje a un blanco distante en sólo un pequeño número de pasos, lo cual indica que hay algo muy especial sobre la estructura de la red.

En una *red aleatoria* (sección 2.3.1), por ejemplo, existen caminos cortos entre nodos, pero nadie sería capaz de encontrarlos a partir de la información que las personas solemos tener en situaciones reales.

Modelos Redes Complejas: Topologías y distribuciones de de grado.

Uno de los principales temas de investigación en redes, es tratar de modelar la estructura topológica de las redes reales encontradas en la naturaleza, a través de un modelo matemático. Lo anterior ha dado paso a muchas propuestas de modelos discutidos ampliamente en la literatura de redes complejas [91, 93, 94]. A continuación se presentan tres de los modelos más relevantes sobre topología de redes, los cuales han pasado a ser considerados verdaderamente como clásicos dentro del estudio de redes

complejas.

En esta sección se presentan los modelos más populares de grafos introducidos para describir los sistemas reales, al menos en cierta medida. Los gráficos son útiles los modelos nulos en la detección de la comunidad, ya que no tienen estructura de la comunidad, para que puedan ser utilizados para las pruebas negativas de algoritmos de agrupamiento.

Redes Aleatorias y Topología de Poisson (Modelo de Erdös-Rényi).

El modelo más clásico de una red, es el propuesto por los matemáticos húngaros Paul Erdös (1913-1996) y Alfréd Rényi (1921-1970) a finales de la década de los 50's [101, 102], el cual caracteriza a las llamadas ***gráficas aleatorias***[1]. En éstas redes se seleccionan un par de nodos y se añade un conexión entre estos de manera aleatoria. Así cada arista colocada al azar, está presente o ausente con igual probabilidad. De está manera hay dos parámetros para el modelo: (i) el número total de nodos $\boldsymbol{n}$, que es fijo; y (ii) la probabilidad $\boldsymbol{p}$ de que dos nodos arbitrarios estén conectados entre si de forma independiente a otros pares.

El modelo se puede pensar como el proceso de hilvanar un conjunto de n botones esparcidos por el suelo inicialmente desconectados ($n \gg 1$) [96, 97]. Se escoge una pareja al azar y se atan con un hilo, después se escoge aleatoriamente otra pareja diferente y se repite el proceso. Se pueden escoger botones que ya estén conectados con otros botones, pero no se pueden escoger dos botones que ya estén conectados entre sí, es decir no se puede atar más de una vez a la misma pareja de botones. Después de repetir el proceso m veces, se habrán establecido m enlaces en un conjunto total de n botones, generando así una red. Sin embargo, la pregunta es entonces ¿cuál es la distribución de conexiones $P(k)$ en la red resultante? [96, 97]

Dado que el número total de parejas que se pueden formar con n nodos es $\frac{n(n-1)}{2}$, entonces la probabilidad p de que dos nodos arbitrarios estén conectados, es el total de las m aristas en la red, entre todas las parejas posibles *i.e.*: $p = \frac{2m}{n(n-1)}$. De ésta manera, un nodo arbitrario en la red i puede en enlazarse a lo más a $n-1$ otros nodos, sin embargo, de los m enlaces totales, no necesariamente todos cuentan con el nodo i. Por lo que podemos suponer que de los m enlaces existentes, el nodo i está solamente en k de ellos. Así, la probabilidad $P(k)$ de que un nodo i pertenezca a una de estas k parejas de las $n-1$ posibles (*i.e.* que cuente con k conexiones) en términos p es:

$$P(k) = \binom{(n-1)}{k} p^k (1-p)^{(n-1)-k}$$

[1] Sin embargo existen propuestas previas formuladas de forma independiente por Solomonoff y Rapoport en 1951 [122]

Por lo que la distribución de grado $P(k)$ es una *distribución binomial* para n y m finitas, donde el numero esperado de aristas es $m = p\left(\frac{n(n-1)}{2}\right)$. Si el número de ensayos en esta distribución binomial tiende a infinito y $z = (n-1)p$ se mantiene constante, entonces $P(k)$ converge a una ***distribución de Poisson***.

Para términos de la red, dado que $p = \frac{2m}{n(n-1)}$, entonces $z = \frac{2m}{n}$. Con lo que si la red es muy grande y tomamos el límite $n \to \infty$ de tal forma que la cantidad z permanezca finita, se obtiene que:

$$P(k) = e^{-z}\frac{z^k}{k!} \tag{2.7}$$

Donde el grado promedio es $\langle \mathbf{k} \rangle = z = \frac{2m}{n} = p(n-1)$, por lo que la mayoría de nodos tienen aproximadamente el mismo grado, con un valor cercano al grado promedio $\langle k \rangle$ de la red. Esta distribución es característica de las *redes aleatorias*, la cual disminuye rápidamente para grados altos y cuenta con un pico en $P(\langle k \rangle)$.

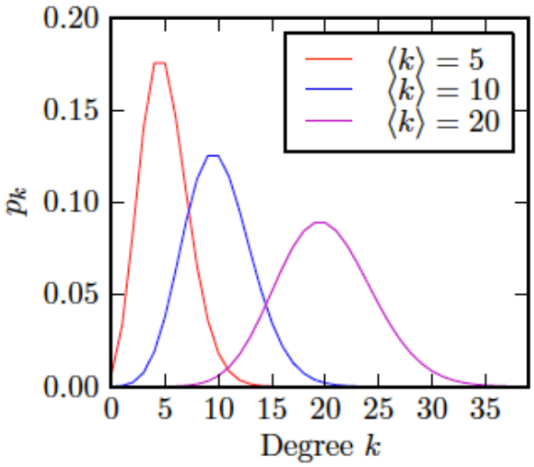


Figura 2.12: **Distribuciones de grado para el Modelo de Erdös-Rényi**. Se muestran tres diferentes distribuciones de grado para redes generadas con el *Modelo de Erdös-Rényi*: $z = \langle \mathbf{k} \rangle = 5$, $\langle \mathbf{k} \rangle = 10$, $z = 20$.

Percolación y la *Isla Gigante*

La propiedad más notable de este tipo de redes, es la *transición de fase* observada mediante la variación de $\langle k \rangle$ en el límite $n \to \infty$. Erdös y Rényi estudiaron cómo cambia la topología de las gráficas aleatorias en función de del numero de aristas promedio $\langle k \rangle$. Cuando $\langle k \rangle$ es pequeña, la gráfica probablemente se fragmente en muchos pequeños

grupos de nodos conectados entre si, llamados *componentes*. A medida que aumenta $\langle k \rangle$, los componentes crecen, en primer lugar por la vinculación a los nodos aislados y más tarde por conexión con otros componentes.

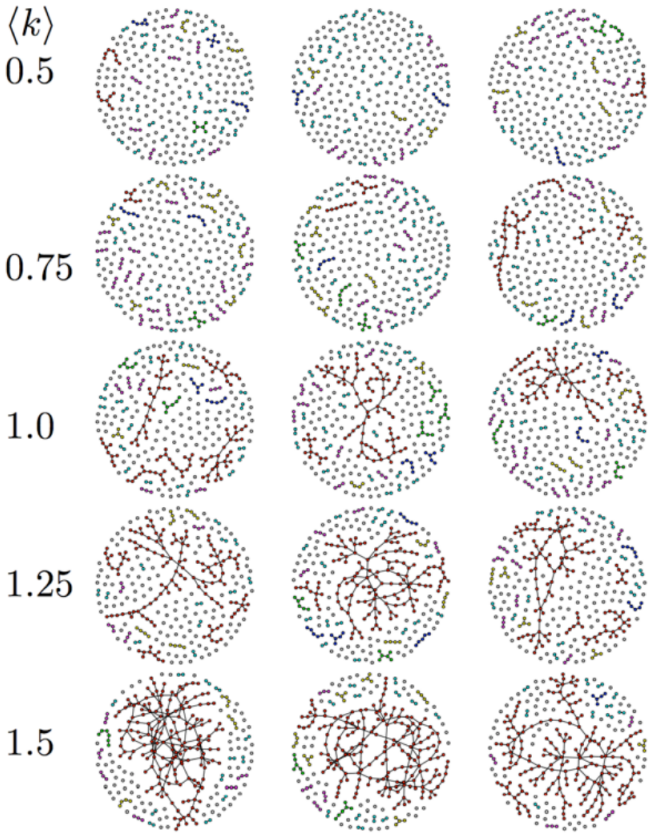


Figura 2.13: **Redes generadas con el *Modelo de Erdös-Rényi***. Se muestran tres diferentes instancias de redes generadas con el *Modelo de Erdös-Rényi* para varios valores de $z = \langle \mathbf{k} \rangle$.

Regresando a la analogía de los botones hilvanados[1], si se levanta lentamente un botón al azar del piso y está ligado a otros botones, ya sea directa o indirectamente, estos serán levantados también. Entonces, ¿qué sucede? ¿Es posible levantar un botón aislado, un pequeño grupo o una malla inmensa? Intuitivamente es claro que si m (el número total de enlaces) es pequeño comparado con n (el número total de botones), entonces la red resultante estará desmembrada en varias *islas* pequeñas. Dentro de cada isla los botones estarán hilvanados entre sí, pero estarán desconectados de las otras islas.

[1]Propuesto por M. Aldana en [96, 97]

Sin embargo, si m es muy grande comparado con n, terminaremos con casi todos los botones hilvanados unos con otros. Probablemente haya islas muy pequeñas desconectadas de la red principal, pero seguramente la gran mayoría de botones formarán parte de una isla principal: la isla gigante [96, 97].

Así, para $\langle k \rangle < 1$ (*i.e.* valores pequeños de p) la red consiste en muchos pequeños componentes. Sin embargo, para $\langle k \rangle < 1$ (p suficientemente grande), uno de los componentes comienza a agrupar a la mayoría de los nodos y aparece en la red el llamado ***componente** (o isla) **gigante***. De ésta manera, ocurre una transición de fase en donde muchos grupos se ínterconectan espontáneamente para formar un único componente en la red. El umbral de percolación donde ocurre dicha transición de fase es $p = \frac{1}{n-1}$, es decir para $\langle k \rangle = 1$ o bien $m = \frac{n}{2}$. Para $m > \frac{n}{2}$, el tamaño de este *componente gigante* escala linealmente con n, mientras que su rival más cercano contiene aproximadamente sólo $log(n)$ nodos. Además de que todos los nodos del componente gigante están conectados entre sí por "*caminos cortos*", donde la longitud máxima del camino más corto entre dos nodos cualesquiera crece lentamente, como $log(n)$ [89, 90, 92, 94, 95].

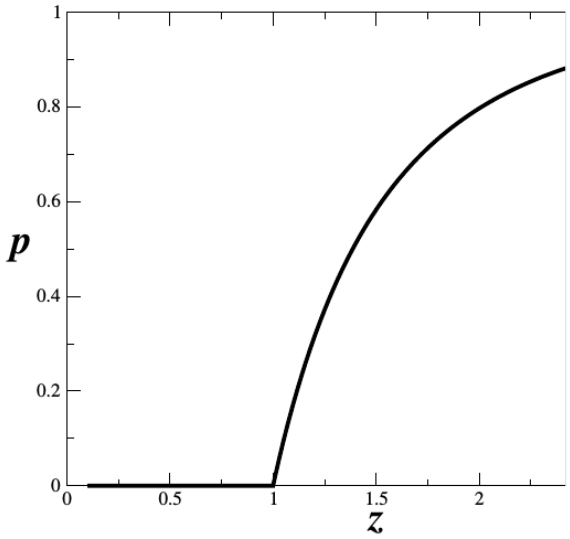


Figura 2.14: **Transición de Fase el *Modelo de Erdös-Rényi***. Se muestran la probabilidad de que todos los nodos de la red pertenezcan al componente conexo más grande (gigante) en función de $z = \langle \mathbf{k} \rangle$. Se muestra que a partir de $z = \langle \mathbf{k} \rangle = 1$ existe una *transición de fase* y el sistema percola hacia un sólo *componente gigante*. Figura tomada de las notas de Maximino Aldana..

Distancia entre nodos y Clustering Coefficient

Distancia promedio entre nodos (*SPL*) La distancia entre dos nodos cualesquiera en una red aleatoria con n nodos es muy pequeña y crece logarítmicamente con n. Es posible estimar aproximadamente como crece la longitud promedio del camino más corto (shortest path length) entre cualesquiera dos nodos $\langle l \rangle$ para el *modelo ER* con n nodos y probabilidad p.

En general si tomamos cualquier nodo arbitrario en la red con k vecinos, el número de nodos total de nodos n^* que están a una distancia l de dicho nodo central crece exponencialmente como $n^* \sim k^l$. Así, dado que en una red aleatoria todos los nodos tienen aproximadamente el mismo numero de vecinos $pn \sim \langle k \rangle$ y que la distancia promedio entre ellos es $\langle l \rangle$, entonces podemos estimar que el número total de nodos n en la red crece como $n \sim \langle k \rangle^{\langle l \rangle}$ [92] y así por lo tanto:

$$\langle l \rangle \sim \frac{log(n)}{log(\langle k \rangle)}$$

es decir, el valor del camino promedio más corto es pequeño incluso para redes muy grandes [89, 90, 92, 94, 95]. Esta *pequeñez* se conoce como *efecto de mundo pequeño* , de la cual hablaremos más adelante.

Clustering Coefficient Por último, el *Clustering Coefficient* esperado para una red aleatoria es p, dado que es la probabilidad de que dos nodos se conecten si son vecinos del mismo nodo o no. Dado que $p = \frac{2m}{n(n-1)}$ y que $\langle k \rangle = \frac{2m}{n}$, entonces para el *modelo ER*:

$$C = p \simeq \frac{\langle k \rangle}{n}$$

sin embargo las redes reales, se caracterizan por valores mucho mayores de Clustering Coefficient en comparación con los redes aleatorias del mismo tamaño.

Crecimiento en redes aleatorias (topología exponencial) El modelo de Erdös y Rényi (*modelo ER*), describe la distribución de vecinos para una red con un **número de nodos fijo**. Sin embargo se han planteado modelos de crecimiento de redes, en el cual el número de nodos aumenta con el tiempo, si un nuevo nodo se enlaza de forma aleatoria como en el *modelo de ER*, la red resultante tendrá lo que se conoce como *Topología de Exponencial* para la distribución de conexiones:

$$P(k) = Ce^{-\alpha k}$$

Este tipo de distribución de grado se a podido observar en redes reales como redes de energía y redes de ferrocarril y con cortes en la red de actores de cine y algunas redes

de colaboración [90].

Finalmente, en las décadas transcurridas desde el trabajo pionero de Erdös y Rényi, las *gráficas aleatorias* se han estudiado profundamente dentro de la matemática pura. También han servido como arquitecturas de acoplamiento ideales para modelos dinámicos de redes genéticas, ecosistemas y para modelar la propagación de enfermedades infecciosas o virus de computadora en la web. Sin embargo, es importante destacar que durante muchos años se pensó que el mecanismo de formación de redes descrito por el *modelo ER*, era adecuado para explicar el origen de ciertas redes reales como las de amistades. Sin embargo, a pesar de la relevancia histórica de este modelo, uno de los hechos más trascendentes en el estudio de las *redes complejas* fue el descubrimiento de que para la mayoría de las grandes redes que hay en la naturaleza, la distribución grado no sigue una distribución de Poisson.

Redes de mundo pequeño (Modelo de Watts y Strogatz).

Como ya mencionamos, el diámetro de un grafo aleatorio con n nodos es pequeño y crece logarítmicamente con el número de nodos n, sin embargo, esta propiedad conocida como *efecto de mundo pequeño* (sección 2.2.2) es muy común en muchas redes reales. Por ejemplo, una de las primeras evidencias de que las redes sociales se caracterizan por caminos de longitud pequeña fue proporcionada por una serie de experimentos llevados a cabo por el famoso psicólogo Stanley Milgram [114] (sección). De la misma manera, el coeficiente de agrupamiento esperado en una red aleatoria es p. Aún así, las redes reales se caracterizan por valores mucho mayores del coeficiente de agrupamiento en comparación con grafos aleatorios del mismo tamaño (mismo numero de nodos).

Con el objetivo de crear un modelo para redes reales, en 1998 Duncan Watts y Steven Strogatz presentaron un modelo muy importante que provocó un enorme interés en la representación en red de sistemas reales, el llamado (*modelo WS*) [115], donde demostraron que el *efecto de mundo pequeño* y un coeficiente de agrupamiento alto pueden coexistir en el mismo sistema. Diseñaron una clase de redes que resultan una interpolación entre una red regular (que tiene un muy alto coeficiente de agrupación), y una red aleatoria que tiene la propiedad de mundo pequeño. Se parte de una red regular de anillo (lattice) en el que cada nodo tiene grado k, y con una probabilidad p_r cada arista se reconecta a un nodo diferente.

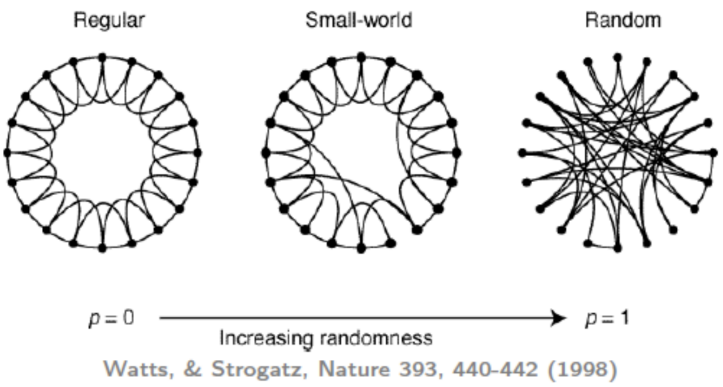


Figura 2.15: ***Modelo de Watts-Strogatz***. Se muestra el procedimiento bajo el cual se construyen redes usando el *Modelo de Watts-Strogatz* al reconectar los nodos con una probabilidad p_r. Al aumentar la probabilidad de reconexión p_r se nota una transición de la red en un régimen regular (*lattice*) hacia una red aleatoria tipo *Erdös-Rényi*.

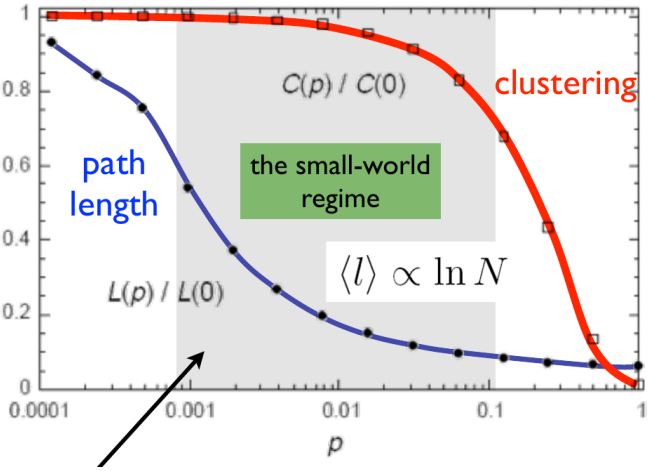


Figura 2.17: **Clustering Coefficient y Shortest Path Length en el *Modelo de Watts-Strogatz***. Se muestran la gráfica de $\langle l \rangle$ y $\langle C \rangle$ vs. la probabilidad de reconexión p_r. La zona gris marcada con una flecha muestra un régimen en el que el *efecto de mundo pequeño* y un coeficiente de agrupamiento alto coexisten en el mismo sistema, lo cual es una característica de las redes reales.

Resulta que valores bajos de p_r son suficientes para reducir considerablemente la

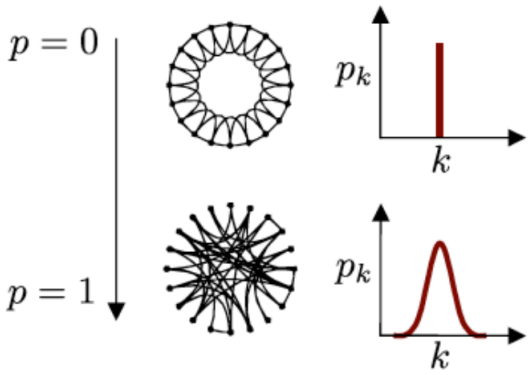


Figura 2.16: **Cambio en la distribución de grado en el *Modelo de Watts-Strogatz***. La distribución de grado pasa de una *delta de Dirac* (todos los nodos con exactamente el mismo grado) a una distribución de una red aleatoria *Erdös-Rényi* (la mayoría de los nodos tienen un grado muy similar, cercano al promedio.

longitud de los caminos más cortos entre nodos, porque las aristas reconectadas actúan como atajos entre las regiones inicialmente alejadas de la red. Por otra parte, el coeficiente de agrupación sigue siendo alto, ya que algunas aristas reconectadas no perturban apreciablemente la estructura local de la red, que sigue siendo similar a la red anillo original. Con $p_r = 1$ todas las aristas son reconectados y la estructura resultante es una red aleatoria de Erdös-Rényi, con topología de Poisson [115].

	path lengths	clustering
Erdös-Renyi networks	short	low
Real-world networks	short	high
Regular networks with triangles	long	high

Figura 2.18: **Comparación de regímenes de redes**. Se muestra los regímenes de redes entre los que se encuentran las redes reales: redes aleatorias (*Modelo de Erdös-Renyí*) y redes ordenadas (*lattices*). A pesar de no ser perfectos tanto los modelos de *Barabasi-Albert* como de *Watts-Strogatz* se acercan a las redes reales.

Crecimiento en redes mediante enlace preferencial y topolgía de Libre de Escala (Modelo de Albert-Barabasi).

El trabajo seminal de Watts y Strogatz provocó un enorme interés hacia la representación en red de sistemas reales. Alrededor de 1998 se comenzaron a publicar los primeros estudios de redes reales a partir de bases de datos [89, 90, 92, 94, 95], en estos se encontró de manera muy amplia que las redes del mundo real son muy diferentes a los *grafos aleatorios* en su distribución de grado, puesto que la forma de $P(k)$ se desvía considerablemente de la distribución de Poisson y que la *topología exponencial* aparece sólo algunas veces en las redes reales.

Así, en la mayoría de las redes estudiadas en la naturaleza la distribución de grado es muy heterogénea, con muchos nodos que tienen pocos vecinos conviviendo con algunos nodos con muchos vecinos. Dicha distribución está altamente sesgada y tiene una *cola larga* a la derecha donde se agrupan la mayoría de los nodos cuyos grados son valores bajos, mientras que a la izquierda hay pocos nodos cuyos valores de grado están muy por encima de la media. En estos casos la cola de la distribución se puede describir con una buena aproximación como una *ley de potencias*[123], donde existen nodos con muy pocas conexiones, nodos medianamente conectados y nodos extremadamente conectados de ahí la expresión ***Redes con Topología Libre de Escala (Scale Free Networks)***[124]. Los nodos altamente conectados se denominan (*hubs*), y así como puede haber nodos con miles de conexiones también hay nodos con una sola conexión. Así, la distribución de grado toma la siguiente forma:

$$P(k) = Ck^{-\gamma} \tag{2.8}$$

Las redes con *topología libre de escala* son muy diferentes estructuralmente a las redes con *topología de Poisson*. Estás últimas son muy homogéneas, mientras que en las primeras la característica más importante es la alta heterogeneidad de grado, la cual es responsable de una serie de características notables, como la ausencia de un umbral de percolación, la resistencia a fallos aleatorios o ataques [89] y se ha mostrado su eficiencia en la propagación de epidemias [125]. Sin duda el resultado más sorprendente de los estudios de redes reales, es la ubicuidad de la topología libre de escala, dado que la ley de potencias para distribuciones de grado se ha observado en redes de citas científicas, el World Wide Web, la Internet, en redes metabólicas, redes de llamadas telefónicas, y redes de contactos sexuales humanos [90].

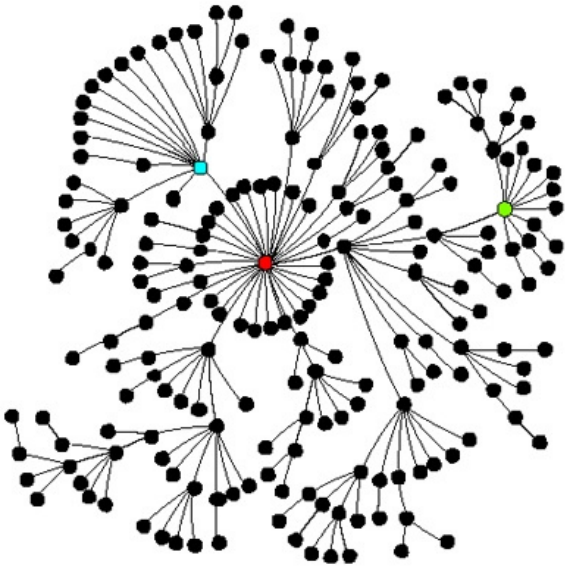


Figura 2.19: **Red Libre de Escala generada con el *Modelo de Barabasi-Albert*.** Nótese la presencia de nodos *hubs* que acaparan muchas conexiones, mientras la mayoría de los nodos sólo tiene pocas conexiones.

Modelo de Barabasi-Albert.

El modelo más popular para generar una red con una distribución de grado que sigue una ley de potencia es el modelo propuesto por Lazlo Barabási y Reka Albert[1] [124] (*modelo BA*), en el cual el número de nodos aumenta con el tiempo y la red se crea con un procedimiento dinámico donde los nodos se añaden uno por uno (modelo de crecimiento de red).

La probabilidad de que un nuevo nodo establezca una conexión con un nodo preexistente es proporcional al grado de este último (enlace preferencial) $p_i = \dfrac{k_i}{\sum_j k_j}$. De esta manera, los nodos con grado alto tienen más probabilidad de ser seleccionados como vecinos por nuevos nodos, y si esto sucede, su grado aumenta aún más por lo que será aún más probable de ser elegido en el futuro. En el límite asintótico del número de nodos ($n(t) \to \infty$), esta estrategia genera una red con una distribución de grado caracterizada por una ley de potencia $P(k) \sim k^{-3}$ (con exponente $\gamma = 3$). En la figura 2.19 se muestra un ejemplo gráfico de la red, sin embargo la información real acerca de ésta se encuentra en su distribución de grado.

[1] Aunque existen modelos anteriores desarrolladas por *Simon* y *de Solla Price* [89, 90].

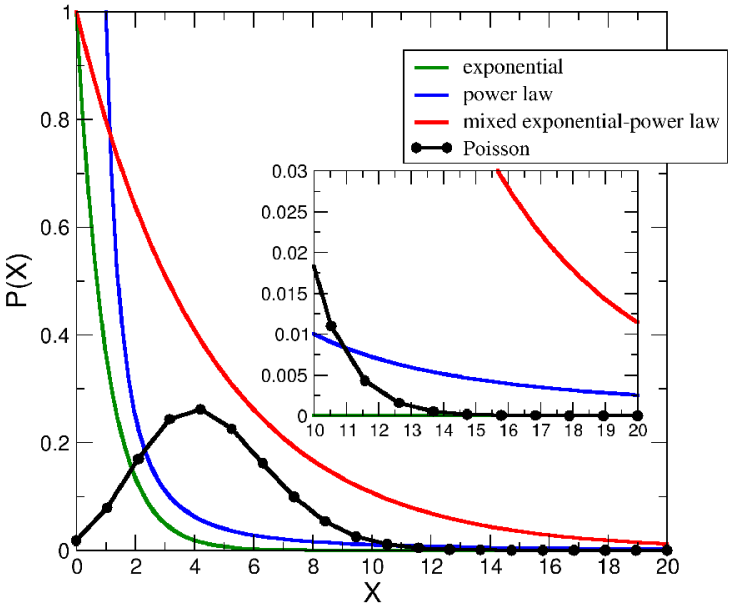


Figura 2.20: **Comparación de distribuciones de grado respecto del *Modelo de Barabasi-Albert***. Se muestran comparaciones con las distribuciones de grado del *Modelo de Erdös-Renyí* (Poisson) y Topología y exponencial.

En general, los decaimientos de ley de potencia de las distribuciones de grado observadas en las redes reales se caracterizan por un intervalo de valores exponentes entre 2 y 3, mientras que el modelo BA produce un valor fijo.

Las redes con distribuciones de grado que siguen una ley de potencia son sorprendentes ya que no se esperaba que existieran, dado que existen muy pocos procesos conocidos que generan distribuciones libres de escala mientras que en la naturaleza existen muchos procesos aleatorios que generan distribuciones de Poisson o distribuciones exponenciales. Dado lo anterior, estas redes han sido el foco de una gran cantidad de atención en la literatura [90, 92].

Distancia entre nodos y Clustering Coefficient Las redes generadas por el *modelo BA* presentan el *efecto de mundo pequeño*, los resultados analíticos muestran que $\langle l \rangle \sim \frac{ln(n)}{ln(ln(n))}$ [126]. Asimismo el Clustering Coefficient en el *modelo BA* decae con el tamaño de la red[89] siguiendo aproximadamente una ley de potencias $C \sim n^{-\frac{3}{4}}$ lo cual es mucho muy pequeño comparado con el de las redes reales, lo que implica que el modelo aún es perfectible y que no es un modelo general al que se puedan ajustar las redes reales.

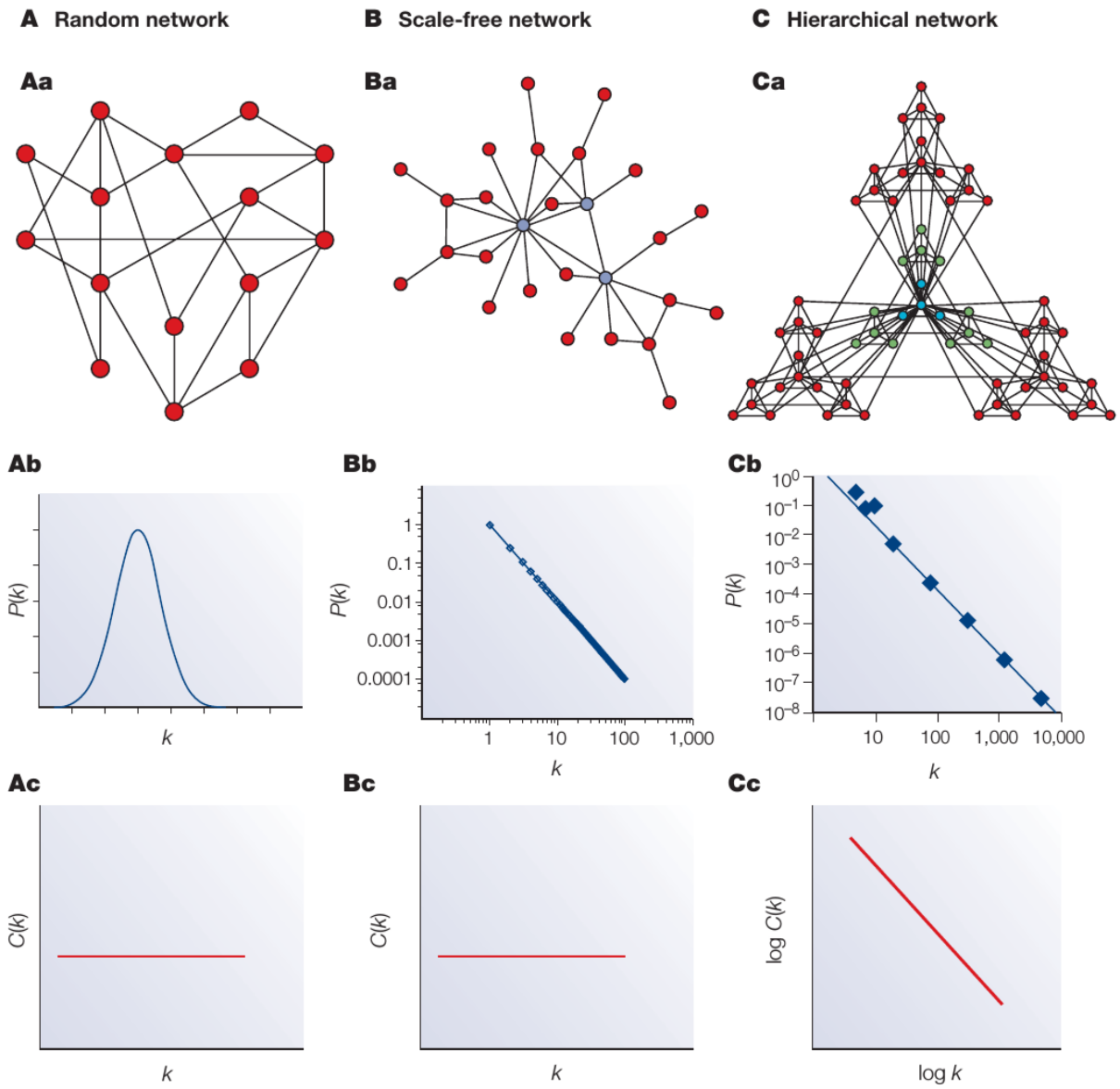


Figura 2.21: **Comparación de Topologías de Red**. Se muestra la diferencia entre 3 principales topologías de red en términos de su distribución de grado y de Clusetering Coefficient.

Finalmente, cabe mencionar que muchos refinamientos del *modelo BA*, así como muchos modelos diferentes se han presentado posteriormente para reflejar lo mejor posible para las características observadas en las redes reales [89, 90, 91, 92, 93, 94, 95, 117].

Estado del Arte en Redes Complejas.

Lo expuesto en las secciones anteriores de este capitulo conforman el cuerpo clásico de estudio de las redes complejas. Las nociones de modularidad que describiremos en el capitulo siguiente y que son el modelo principal del estudio que se presenta en este texto, tienen su fundamento en este marco teórico clásico. Sin embargo el estudio de las redes complejas ha seguido siendo un campo muy activo dando paso a lo que hoy conocemos como *Ciencia de Redes*.

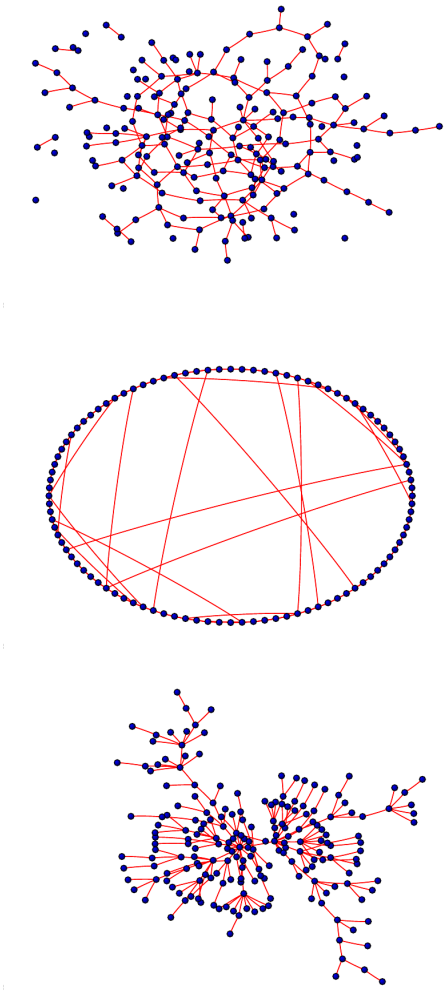


Figura 2.22: **Comparación de Modelos de Red**. Arriba: una red aleatoria generada con el *Modelo ER*, en medio: una red aleatoria generada con el *Modelo WS* y abajo: una red aleatoria generada con el *Modelo BA*.

El estudio de propiedades en redes complejas así como la propuesta de nuevos enfoques en modelos topológicos han puesto en marcha un renacimiento del modelado de redes en los últimos años, sin embargo los modelos clásicos expuestos (*ER*, *BA* y *WS*) siguen siendo ampliamente usados en muchos campos y sirven como punto de referencia para muchos estudios empíricos, por otra parte se han planteando variantes de estos modelos dando lugar al estudio de tres clases principales de paradigmas de modelado. Las redes aleatorios (variantes del modelo ER), otros que interpolan entre redes regulares y redes aleatorias así como varios modelos que reproducen la topolgía de libre de escala y se centran en las propiedades dinámicas de las redes las cuales merecen una revisión en si [89, 90, 91, 92, 93, 94, 95, 117].

Asimismo, además de las características topológicas mencionadas, existen muchas otras muy bien estudiadas en otros tipos de redes como las bipartitas o bien otras cuya distribución de grado puede ser más complicada. Para las redes bipartitas, por ejemplo, hay dos distribuciones de grado, uno para cada tipo de nodo. Para redes dirigidas cada nodo tiene tanto el grado de entrada como el grado de salida, y la distribución de grado por lo tanto, se convierte en un función de dos variables. Por ejemplo en los estudios empíricos de la Web, se muestran por lo general sólo las distribuciones individuales de grado de entrada y de salida, sin embargo, se ha encontrado que los grados de entrada y salida están fuertemente correlacionados en algunas redes, lo que sugiere que se puede obtener más información de la distribución conjunta de lo que normalmente se aprecia [90].

Otras aportaciones más recientes a la teoría de redes (provenientes en general de la física estadistica) es por ejemplo el estudio de la ***entropía* en redes** por Bianconi et.al. [127, 128], sus aportes han tenido una aplicación importante en redes biológicas mostrando formalmente que las diferencias en la expresión genética entre tejido normal y sano están anticorrelacionadas con cambios en la entropía local de la red [88]. Así también el estudio de la llamada **matriz *nonbacktracking***, está matriz es usada para identificar *caminos sin regresos*, es decir, contiene la información de a que nodos se puede acceder desde un nodo j una vez que se ha llegado a este desde el nodo i pero sin poder regresar a este último [129]. La *matriz nonbacktracking* ha tenido aplicaciones sobre *resistencia en la red* y asimismo es el elemento teórico principal de una serie de nuevos algoritmos que tienen el objetivo de redimir la teoría espectral de redes para detección de comunidades (véase sección 3.3.3.3 del siguiente capitulo) [130].

Asimismo el espectro de la *matriz nonbacktracking*[131] ha sido usado recientemente por Morone y Makse, quienes mapean el problema percolación óptima en redes aleatorias para encontrar un conjunto de nodos estructurales llamados *influencers* que propagan información a toda una red [132], estás ideas pueden ser utilizadas, por ejemplo, para demolición (desconexión) optima en redes [133]. Asimismo el *Collective Influence Method* (CI) derivado de estás ha sido aplicado por Teng et.al. para localizar nodos influyentes que maximizan el flujo de información en redes reales como Facebook y Twitter entre

otras, asimismo compararon el CI con otros métodos [134]. Karrer, Newman y Zdeborová mostraron que la percolación en redes tipo árbol puede reformularse como un proceso de paso de mensajes y el uso el inverso del *eigenvector* principal de la *matriz nonbacktracking* les permitió resolver propiedades de percolación en dichas redes, probando su método tanto en redes generadas computacionalmente como en redes reales obteniendo buenos resultados [135].

Redes Multicapa (Multilayer Networks).

Finalmente, a partir de 2014 M. De Domenico, M. Porter y A. Arenas entre otros autores presentaron un novedoso enfoque en el cual se plantean redes compuestas de varias capas conocidas hoy en día como ***Redes Multicapa*** *(Multilayer Networks)*[136, 137, 138]. En este nuevo tipo de modelo plantea varias redes conectadas entre si en varias capas diferentes en donde los nodos no sólo están interconectados entre si en su propia capa, sino también pueden estar conectados con otros tipos nodos de las redes en otras capas (*enfoque Multicapa*) o bien con sigo mismos como la misma entidad en diferentes capas (*enfoque* ***Multiplex***), pensemos por ejemplo en un usuario en las redes de *Facebook* y *Twitter*, es el mismo nodo sólo que en dos redes (capas) diferentes.

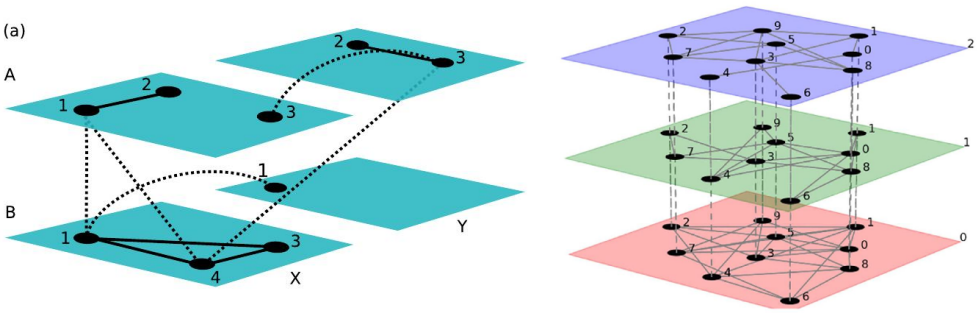


Figura 2.23: **Ejemplos de Redes Multicapa**. Derecha: una *red Multilayer* (los nodos tienen conexiones entre ellos en una capa y entre los de las otras capas). Izquierda: una *red Multiplex* (los nodos son las mismas entidades por lo que están conectados con ellos mismos en las diferentes capas, pero tienen diferentes interacciones con los demás nodos en cada capa).

Este nuevo enfoque ha generalizado las *Redes Complejas* y las Bipartitas bajo un mismo enfoque replanteando la teoría en términos **tensoriales**, *i.e.* una matriz de matrices de adyacencia, una por cada capa[136]. Este enfoque ha sido muy exitoso para modelar *sistemas complejos multiescala* como por ejemplo la regulación transcripcional que ocurre en diferentes capas la *genética*, la de *ARN* y la de *proteínas* donde un nodo

(gen) es una misma entidad en las tres capas y puede interactuar con otro tipo de nodos (figura 2.24). Así entonces, gran parte de la literatura reciente en *fundamentos* de la *Ciencia de Redes* se enfocado a generalizar las propiedades estructurales y dinámicas de las redes complejas a las *redes multicapa*[139, 140] así como visualizarlas[141]. Con lo que se han logrado técnicas para encontrar nodos centrales y relevantes[142], así como generalizar la teoría de entropía[143] y percolación[144] para este nuevo tipo de redes.

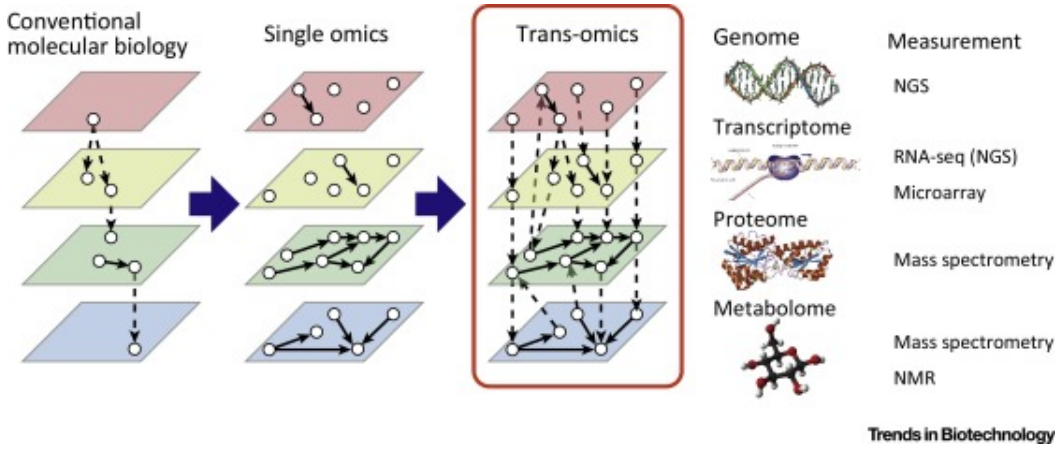


Figura 2.24: Las Redes Multicapa parecen ser un marco teórico prometedor para modelar redes de sistemas biologicos moleculares a diferentes escalas.

Capítulo 3

Modularidad y métodos computacionales de detección de comunidades en redes complejas.

Como vimos en el capitulo anterior las Redes Complejas son modelos teóricos útiles para representar Sistemas Complejos los cuales en general presentan una estructura modular y jerárquica en su organización. Así, las redes que modelan Sistemas Complejos (como los biológicos) no son *gráficas aleatorias*[1], en éstas hay grandes heterogeneidades que revelan un alto grado de orden y organización. La distribución de grado a menudo sigue a una ley de potencia y por lo tanto, muchos nodos con bajo grado coexisten con algunos nodos con alto grado. Por otra parte, la distribución de aristas no es sólo globalmente, sino también, localmente heterogénea, con altas concentraciones de aristas dentro de grupos especiales de nodos, y bajas concentraciones entre estos grupos dentro de un mismo componente conexo de la red. Estos patrones de organización global de grandes redes complejas implican la presencia de subunidades estructurales (subredes) llamadas *módulos* o *comunidades*[2].

Las ideas de este capitulo están basadas muy fuertemente en las revisiones de Santo Fortunato *et al.* sobre detección de comunidades en redes complejas [146, 147]

[1] Aunque, la modularidad también se ha estudiado en redes aleatorias clásicas (tipo ER) [145]

[2] El termino *comunidad* se adoptó de las ciencias sociales con la intensión de no confundir con el concepto de *clustering*.

Modularidad y estructura modular en Redes Complejas.

Las redes complejas pueden rescatar la organización modular y jerárquica de los sistemas complejos bajo una de sus características topológicas de gran escala más importantes; la llamada **estructura modular** (o bien estructura de comunidad), ***community strucuture***, la cual refleja la organización modular de la red completa y que es el tema del presente capítulo. En la figura 3.1 se muestra un ejemplo esquemático muy simple de una pequeña gráfica con estructura de comunidad (o modular).

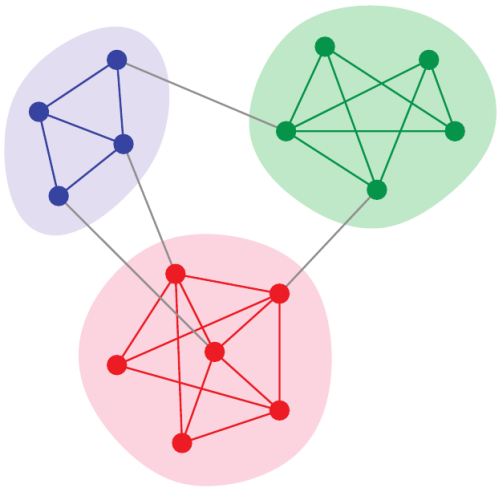


Figura 3.1: **Ejemplo esquemático simple de una pequeña gráfica con estructura modular**. Se muestran tres conjuntos de nodos (rojo, verde y azul) con una mayor densidad de aristas entre ellos respecto a los otros conjuntos de nodos.

Como veremos mas adelante (sección 3.1.2), las comunidades también están presentes en muchas redes biológicas, por ejemplo, en las redes de interacción proteína-proteína las comunidades pueden ser probablemente grupos de proteínas que tiene la misma función específica dentro de la célula o en redes metabólicas pueden estar relacionadas con módulos funcionales. Así, en este capitulo discutiremos la relevancia de la modularidad en redes complejas y cómo es que se aborda ésta característica matemáticamente.

Ejemplos de Redes reales con estructura modular.

Muchas redes reales muestran *estructura modular*, por ejemplo las redes sociales ofrecen una amplia variedad de organizaciones de grupos posibles: familia, trabajo y amistades, aldeas, pueblos, naciones, así como la creación de grupos virtuales y comuni-

dades en línea que viven en la World Wide Web, donde además las comunidades pueden corresponder a grupos de páginas que se ocupan de los mismos temas o relacionados. Asismimo, la organización modular también se produce en muchos sistemas informáticos, económicos, políticos, etc. [146]. En esta sección vamos a presentar algunos ejemplos estudiados de redes reales (no biológicas) con estructura de comunidad.

Modularidad en Redes Sociales.

Las redes que representan las interacciones sociales entre personas se han estudiado durante décadas y son ejemplos paradigmáticos que ponen de manifiesto la *estructura modular*, es decir, grupos de nodos que tienen una alta densidad de aristas dentro de ellos, con una menor densidad de aristas entre los grupos. La misma palabra comunidad se refiere a un contexto social. Es común que la gente naturalmente tienda a formar grupos, dentro de sus entorno de trabajo, familia, amigos, a lo largo de líneas de interés, ocupación, edad, y así sucesivamente. Por lo que las comunidades en redes sociales pueden ser círculos de amistad, grupos de personas que comparten intereses comunes y/o actividades, etc.

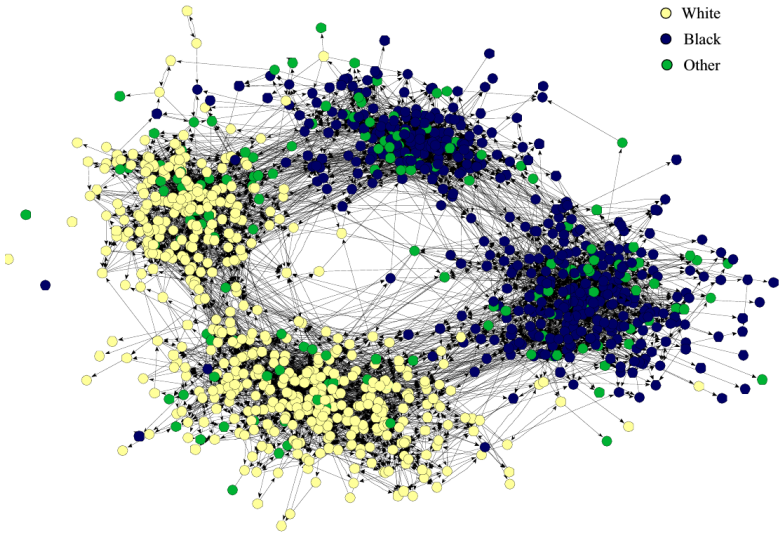


Figura 3.2: **Red de amistades de niños en una escuela de Estados Unidos tomada de un estudio de Moody**. Se puede notar que una de las divisiones de la red es la etnia y otra la división entre la escuela media y la secundaria.

En la figura 3.2 se muestra una visualización de la red de amistades de niños en una escuela de Estados Unidos tomada de un estudio de Moody [148]. Cuando Moody coloreo los nodos de acuerdo con la etnia de los individuos que representan, como se

muestra en la figura, se hace evidente que una de las divisiones principales de la red es la etnia de los individuos, y esto es presumiblemente lo que está impulsando la la formación de las comunidades en este caso, otra división principal visible en la figura es entre la escuela media y la secundaria, que son las divisiones por edad en el sistema educativo estadounidense.

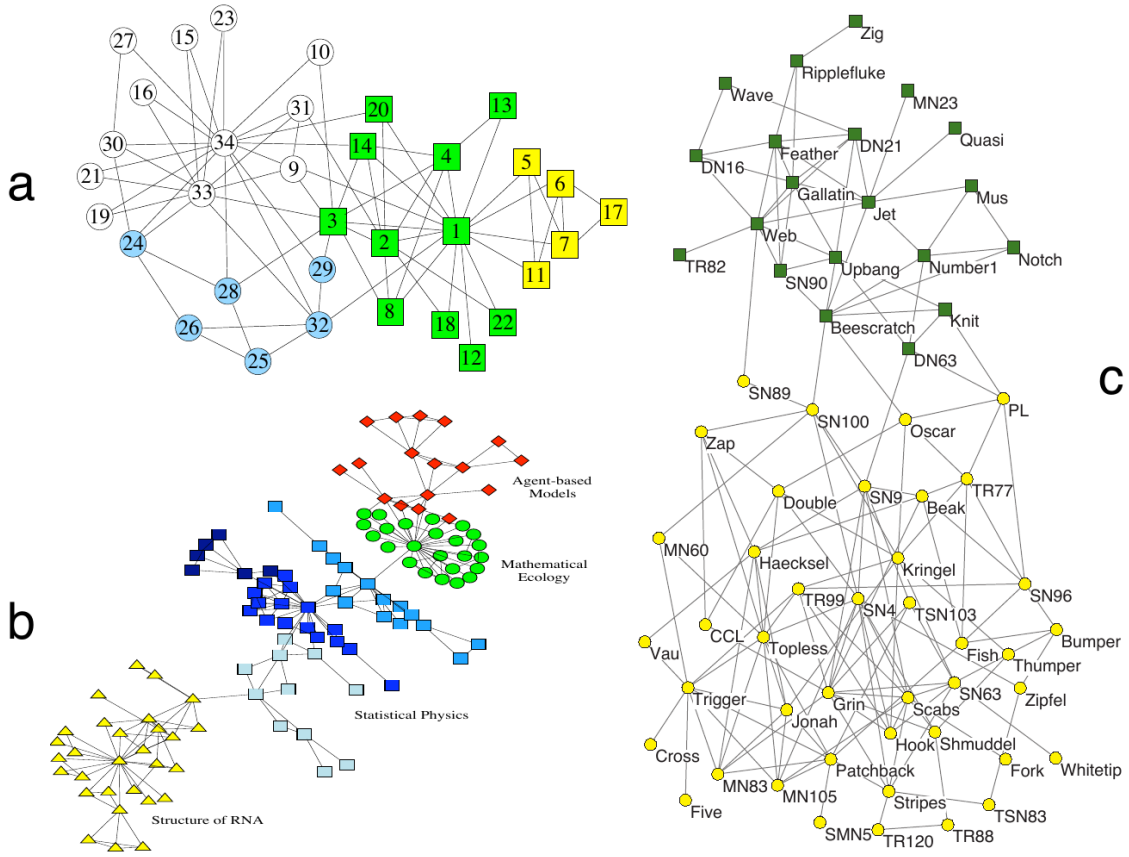


Figura 3.3: **Ejemplos de modularidad en diferentes redes sociales**. **a**: la famosa red del *Club de Katate de Zachary* 3.4.1.1 con una partición en 4 grupos (diferente a la tradicional; **b**: Red de colaboraciones en el *Santa Fe Institute* **c**: Red de delfines nariz de botella de Lusseau.

En la figura 3.3 se muestran algunos ejemplos de redes sociales. El primer ejemplo (3.3 panel **a**) es la red de Zachary de miembros del club de karate [149], una red usada regularmente como referencia para poner a prueba algoritmos de detección de comunidades (sección 3.4.1.1). Se compone de 34 nodos, que son los miembros de un club de karate en los Estados Unidos, que fueron observados durante un período de tres

años. Las aristas conectan individuos que fueron observados interactuar fuera de las actividades del club. En algún momento, un conflicto entre el presidente del club y el instructor dio lugar a la fisión del club en dos grupos separados, apoyando al instructor y al presidente, respectivamente (marcados por cuadrados y círculos). La cuestión es si a partir de la estructura original de la red es posible deducir la composición de los dos grupos. De hecho, al ver la figura 3.3 panel **a** se pueden distinguir dos agrupaciones, una en torno a los nodos 33 y 34 (34 es el presidente), y la otra en torno a 1 (el instructor). También se pueden identificar varios nodos situados entre las dos estructuras principales, como 3, 9, 10; tales nodos son a menudo mal clasificados por métodos de detección comunidades (sección 3.3).

En el panel 3.3 **c** se muestra la *red de delfines nariz de botella* que viven en Doubtful Sound (Nueva Zelanda) analizados por D. Lusseau [150] (sección 3.4.1.2). Hay 62 delfines y las aristas se establecieron entre los animales que se veían juntos más a menudo de lo esperado. Los delfines se separaron en dos grupos después de que un delfín abandonó el lugar por algún tiempo (cuadrados y círculos en la figura). Estos grupos están bastante cohesionados, con varios *cliques*[1] internos y fácilmente identificables: sólo seis aristas unen a los nodos de los diferentes grupos. Debido a que existe una clasificación "natural" tanto para la *red delfines de Lusseau*, como para la del *club de karate de Zachary*, a menudo se utilizan para probar algoritmos de detección de comunidades (ver sección 3.4.1).

Otro ejemplo bien estudiado, son las redes legislativas, que permiten deducir una asociación entre políticos a través de su actividad parlamentaria, la cual puede estar relacionada o no a afiliación partidaria. M.A. Porter *et al.* han llevado a cabo numerosos estudios sobre el tema [151, 152, 153], mediante el uso datos sobre el Congreso de los Estados Unidos. Finalmente, Espinal *et al.* [154] detectaron comunidades usando *Infomap* (sección 3.3.5.4) en una red de personajes de narcotráfico en México construida a partir de minería de texto sobre un libro periodístico. Las comunidades relacionan a personajes de la política mexicana con capos de carteles del narcotráfico mexicano.

Modularidad en Redes de colaboración y citas de artículos.

Las redes de colaboración, en el que los individuos se vinculan si están (o estuvieron) participado en una actividad común, han sido muy estudiadas. Una colaboración, es una prueba de una relación social entre los individuos. En particular el análisis de la estructura modular de redes de colaboración científica ha sido investigado por muchos autores [155] y ha ejercido una gran influencia en el desarrollo de la *ciencia de redes*. La colaboración científica se asocia a coautoría: dos científicos están vinculadas si han sido couautores de al menos un articulo. La información sobre coautorias se puede extraer de diferentes bases de datos de trabajos de investigación. Los módulos indican grupos de personas con intereses comunes de investigación, *i.e.* grupos temáticos o disciplinarias.

[1] Sub-redes totalmente conectadas (*completas*).

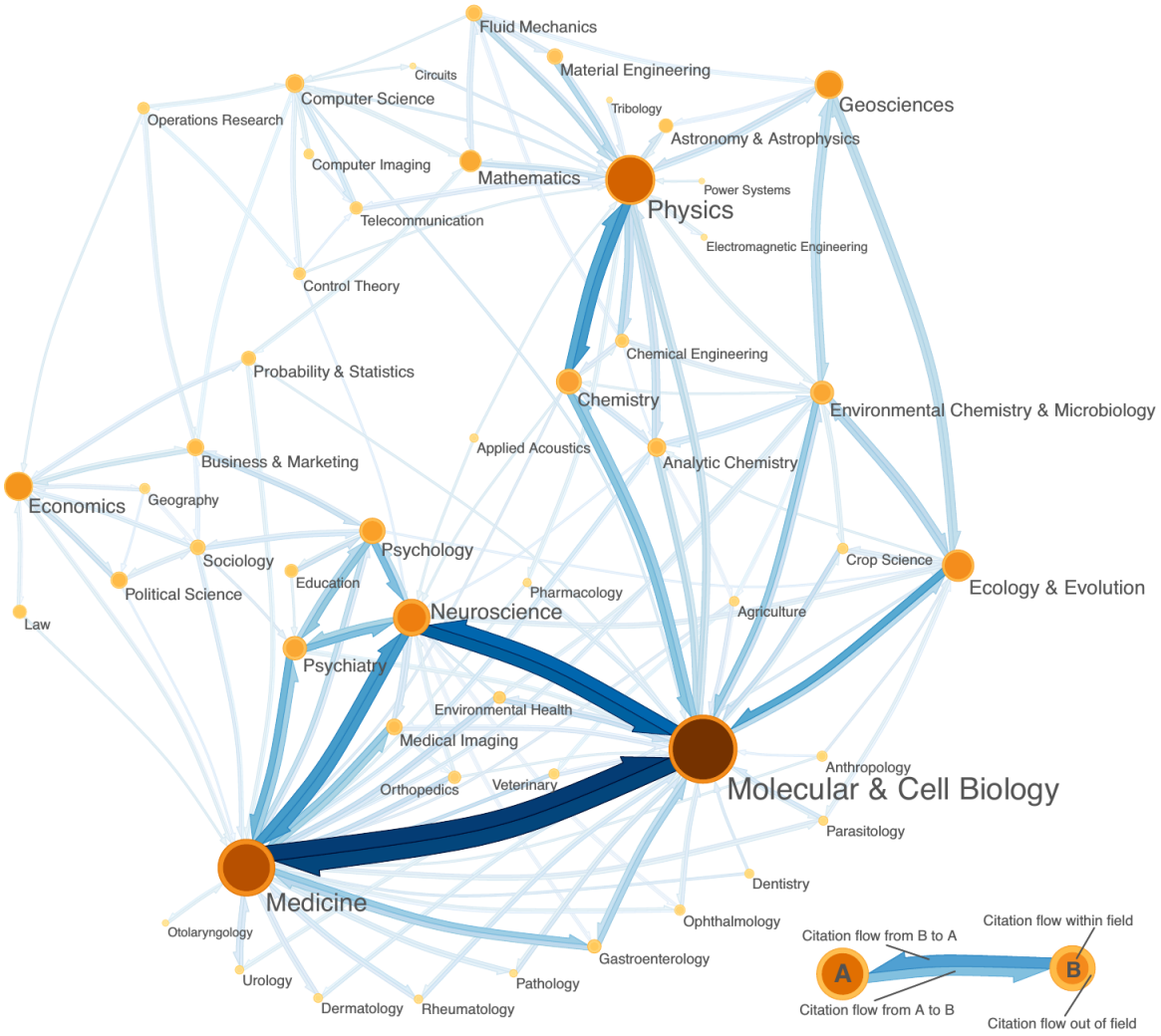


Figura 3.4: **Modularidad en redes de citas**. Mapa de la ciencia de Rosvall y Bergstrom derivado de un análisis de una red de citas que comprende más de 6000 revistas. Los autores usaron su método *Infomap*, el cual cómo veremos más adelante (secciones 3.3.5.4,3.4.4 y 4.1.1) es una de las mejores propuestas para encontrar la estructura modular de una red. Los nodos representan *áreas de conocimiento* (módulos) y las aristas y su grosor el flujo de información (de citación) entre las áreas.

Rosvall y Bergstrom [156] utilizaron una red de citas de más de 6000 revistas científicas para obtener un mapa de la ciencia, figura 3.4. En el trabajo seminal de Girvan y Newman de 2002 [157] (sección 3.3.2.1), los autores aplicaron su método en una red de

colaboración de científicos que trabajan en el *Instituto Santa Fe* (**SFI**), y fueron capaces de discriminar entre las divisiones de investigación. La figura 3.3 **b**, muestra el mayor componente conectado la red, en está hay 118 nodos, que representan a los científicos residentes en el *SFI* y sus colaboradores. Las aristas se colocan entre los científicos que han publicado al menos un artículo juntos. La visualización de la disposición permite distinguir grupos disciplinarios. En esta red se observan muchos *cliques*, dado que los autores del mismo artículo están vinculados entre sí. Asismimo, se nota que no hay más que unas pocas conexiones entre la mayoría de los grupos.

Un estudio de otro tipo de red de colaboración es la de colaboración de músicos de jazz de Gleiser y Danon [158]. Los nodos son músicos, conectados si es tocaban en la misma banda, o bien bandas conectadas si tienen un músico en común. Al aplicar el *algoritmo de Girvan y Newman* (sección 3.3.2.1) se encontró que las comunidades reflejan tanto la segregación racial (con dos grupos etnicos principales que comprenden solamente los músicos de afrodescendientes o anglosajones) como la separación geográfica, debido a los lugares de grabación diferentes.

Modularidad en Redes Informáticas.

Recientemente, las tecnologías de la información y la comunicación (*TICs*) han abierto nuevas formas de interacción entre los individuos, como las comunicaciones de telefonía móvil y las interacciones en línea posibles gracias al Internet. Estos nuevos intercambios sociales pueden ser adecuadamente monitoreados para sistemas muy grandes, incluyendo millones de personas, cuyo estudio representa una enorme oportunidad para la ciencia social.

Reichardt y Bornholdt [159] realizaron un análisis en una red construida a partir de datos sobre ofertas tomadas de la versión alemana de *Ebay* (www.ebay.de). Los nodos son los postores y están conectados si han expresado su interés por el mismo artículo. Las comunidades fueron detectadas con un método desarrollado por los propios autores [160, 161] (sección 3.3.5.1). A pesar de la variedad de artículos que se pueden adquirir a través de *Ebay*, el 85 % de los postores fueron clasificados en unos pocos grandes grupos, lo que refleja amplias categorías de intereses de postores.

Blondel *et al.* analizaron una red de comunicaciones de telefonía móvil entre usuarios de un operador de telefonía belga [162]. Los nodos de la red son 2.6 millones y las aristas están pesadas por el tiempo acumulado de llamadas entre los usuarios en el marco de tiempo de observación. El análisis realizado con una técnica desarrollad por los autores [162](sección 3.3.4.2), ofrece seis niveles jerárquicos. El nivel más alto se compone de 261 grupos con más de 100 nodos, los cuales están claramente dispuestos en dos grupos principales, lingüísticamente homogéneos, lo que refleja la división lingüística de la población belga.

Los sitios de redes sociales, como *Twitter*, *Facebook*, etc., se han vuelto muy popular en los últimos años. Traud *et al.* [163] usaron datos anónimos de Facebook para crear redes de amistad entre estudiantes de diferentes universidades americanas, donde los nodos (estudiantes) están conectados si son amigos en Facebook. Las comunidades se detectaron mediante la aplicación de una variante de la optimización espectral de modularidad de Newman [164, 165] (sección 3.3.4.4). Uno de los objetivos del estudio era inferir relaciones entre las vidas en línea y fuera de línea de los estudiantes. Mediante el uso de información demográfica sobre las poblaciones de los estudiantes, se observa que las comunidades están organizados por año de la clase o por afiliación de casa (dormitorio), dependiendo de la universidad.

Modularidad en Redes biológicas.

Las redes biológicas se caracterizan por una organización modular notable [73, 166] (ver sección 1.3.1 capitulo 1) y como veremos en las sección 3.3 existen muchos métodos para encontrar sus módulos o comunidades[167, 168]. Dicha organización refleja asociaciones funcionales entre sus componentes, por ejemplo, las proteínas tienden a asociarse en dos tipos de módulos celulares: complejos de proteínas y módulos funcionales. Un complejo de proteínas es un grupo de proteínas que interactúan mutuamente en el mismo tiempo y espacio, formando una especie de objeto físico. Ejemplos de ello son los complejos de factores de transcripción, el transporte de proteínas y complejos de exportación, etc., los módulos funcionales en cambio, son grupos de proteínas que tienen lugar en el mismo proceso celular, incluso si las interacciones ocurren en diferentes momentos y lugares. Ejemplos son el módulo de *CDK/ciclina*, responsable de la progresión del ciclo celular, la ruta de respuesta de la feromona de levadura, etc. La identificación de módulos celulares es fundamental para descubrir la estructura y la dinámica de las funciones celulares [169]. Sin embargo, la información sobre las unidades celulares (por ejemplo, proteínas, genes) y sus interacciones es a menudo incompleta o incorrecta, debido al ruido en los datos producidos por los experimentos. Por lo tanto, inferir módulos a partir de la topología de redes celulares permite restringir el conjunto de posibles escenarios y puede ser una guía segura para futuros experimentos.

El estudio de módulos en redes biológicas tiende a irse refinando y consolidando incorporado el reciente enfoque multicapa [170, 171]. Asimismo, estudios recientes en la denominada Medicina de red (*Network Medicine*) (ver sección 1.3.2 capitulo 1) han demostrado que genes y proteínas asociadas a una enfermedad tienden a agruparse en la red (módulo de la enfermedad), lo que que representa una sub-red conectada dentro de un interactoma para una enfermedad determinada [81, 172] (sección 1.3.2). Con lo que se han generando algoritmos para identificar módulos de enfermedad alrededor de un conjunto de proteínas de enfermedades conocidas [80] (sección 1.3.2). A continuación presentaremos algunos ejemplos de modularidad en redes biológicas.

Redes de Interacción de Proteínas.

Las redes de Interacción Proteína-Proteína (*PPI*) son objeto de intensas investigaciones en biología y bioinformática (sección 1.1.1), pues las interacciones entre proteínas son fundamentales para cada proceso en la célula [173]. La figura 3.5 ilustra una red de interacción proteína-proteína (*PPI*) de proteoma de rata [174]. Cada interacción se deriva por homología de las interacciones observadas experimentalmente en otros organismos. En este ejemplo, las proteínas interactúan muy frecuentemente entre sí, ya que pertenecen a células metastásicas, que tienen una alta motilidad e invasividad con respecto a las células normales. Las comunidades fueron detectadas con el *Método de Percolación de Clique* de Palla *et al.* [175] (sección 3.3.6.3) y corresponden a *grupos funcionales*, *i.e.* a proteínas que tienen las mismas funciones o similares, las cuales se espera que participen en los mismos procesos. Los módulos están etiquetados por la función global o la clase de proteínas dominante. La mayoría de las comunidades están asociadas con el cáncer y la metástasis, lo que indirectamente demuestra cuan importante es la detección de módulos en las redes de interacción de Proteína-Proteína.

La organización modular de la red de interacciones de proteína en la levadura (*Saccharomyces cerevisiae*) ha sido motivo de mucha atención en la literatura (sección 1.1.1), Han *et al.* [176] evidenciaron modularidad organizada dinámicamente en esta red. Asimismo Rives and Galitski [177] estudiaron la organización modular de un subconjunto de la red que consta de proteínas (de señalización) que participan en procesos que conducen al microorganismo a una forma filamentosa. Los módulos se detectaron con una técnica de *agrupación jerárquica* (ver sección 3.3.1). Las proteínas que interactúan principalmente con miembros de su propio grupo a menudo son proteínas esenciales, los aristas entre módulos son puntos importantes de comunicación. También para una red de interacciones de proteína de levadura, Palla *et al.* [175] encontraron módulos, pero en este caso sobrelapados, usando su método basado en el análisis de *sobreplapamiento de cliques* en la red, el cual se basa en localizar todos los *cliques* en la red e identificar las comunidades llevando a cabo un análisis de en la matriz de sobrelapamiento (sección 3.3.6.3).

Spirin y Mirny [178] identificaron complejos de proteínas y módulos funcionales en levadura con diferentes técnicas. Se estimó la importancia estadística de los grupos mediante el cálculo de *p-values* al observar a esos grupos en *redes aleatorias* (sección 2.3.1) con la misma secuencia esperada de grado que la red original. De las formas funcionales conocidas los genes de levadura se puede ver que los módulos son normalmente grupos de proteínas con la mismas o consistentes funciones biológicas. De hecho, en muchos casos, los módulos coinciden exactamente con complejos de proteínas conocidas. Los resultados aparecen robustos si se introduce ruido en el sistema, (para simular el ruido presente en los datos experimentales).

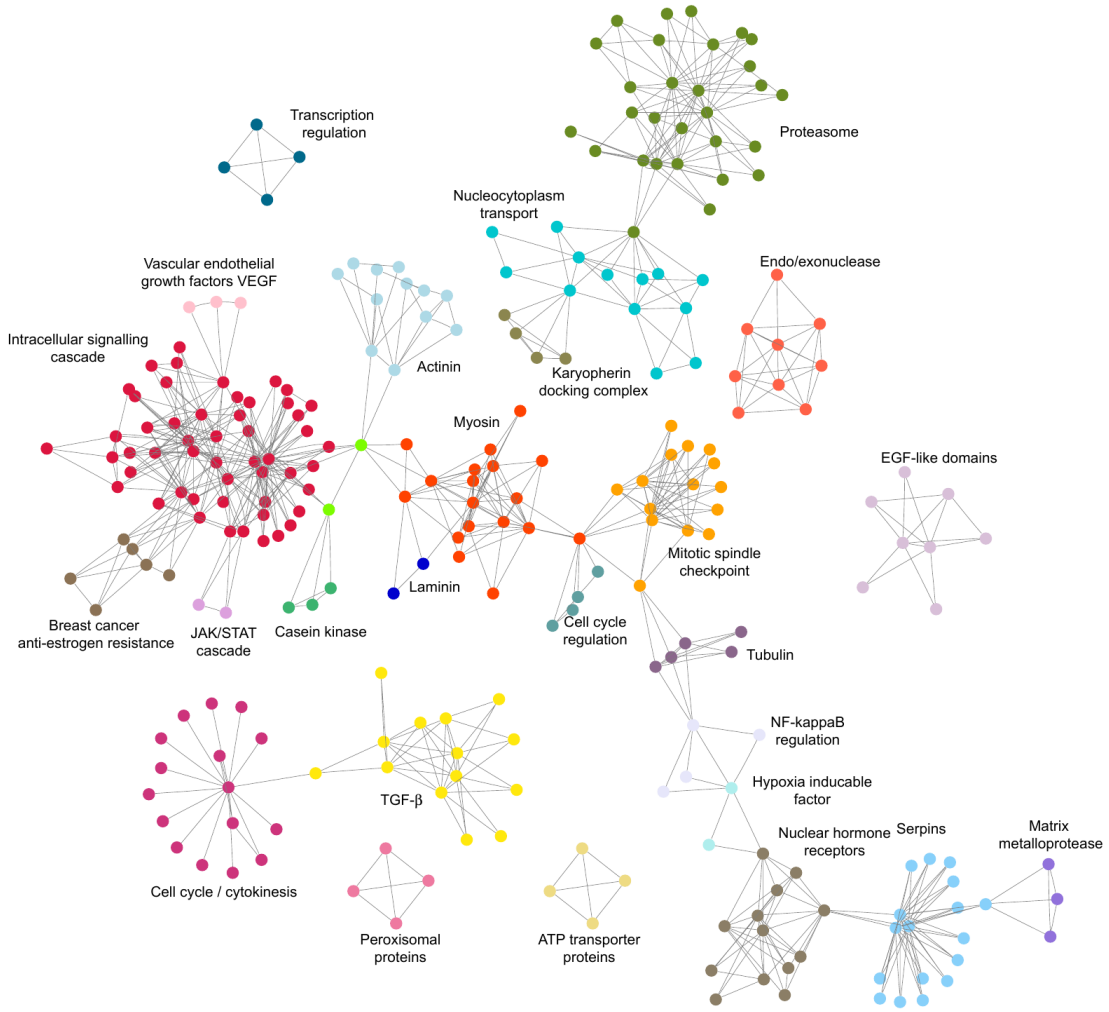


Figura 3.5: **Estructura modular en una red de interacción proteína-proteína**. Se muestran las interacciones entre proteínas en células cancerosas de rata. Las comunidades, etiquetadas por colores, fueron detectadas con el *Método de Percolación de Clique* de Palla *et al.* (sección 3.3.6.3).

Módulos funcionales en la levadura (sección 1.1.1) también se encontraron por Chen y Yuan [179], quienes aplicaron el algoritmo de Girvan y Newman [157] con una definición modificada de intermediación de arista (sección 3.3.2.1). Asimismo, el algoritmo estándar de Girvan-Newman [157] ya había demostrado ser fiable para detectar módulos funcionales en las Redes de Interacción de Proteinas [180]. La novedad de la obra de Chen y Yuan es que se centra en las *Redes de Interacción de Pro-*

teinas ponderadas, donde los pesos proceden de la información derivada a través de *perfiles de expresión de microarreglos* (ver sección 1.2.1 capitulo 1) . Los pesos agregan información sobre el sistema y podrían dar lugar a una *estructura modular* más fiable. Fenotipos similares aparecieron por anulación de genes en el mismo grupo estructural, lo que sugiere que los genes tienen similares funciones biológicas. Por otra parte, los módulos a menudo contenían complejos proteicos conocidos, ya sea totalmente o en gran parte. Finalmente, Chen y Yuan fueron capaces de hacer predicciones de la función desconocida de algunos genes, basándose en el módulo estructural al que pertenecen: la predicción de la función de genes es el resultado más prometedor que se deriva de la aplicación de técnicas de detección de comunidades a las *Redes de Interacción de Proteinas* .

Farutin *et al.* [181] han adoptado un concepto local de comunidad y derivado una *descomposición jerárquica* de *redes de interacción de proteínas* **PIN**, ya que los módulos identificados en algún nivel se convierten en los nodos de una red en el nivel superior. Las comunidades están superpuestas (sección 3.2.4.3), para tener en cuenta el hecho de que las proteínas (y módulos enteros) puede tener diversas funciones biológicas. Las estructuras de alto nivel detectadas en una *PIN* humana corresponden a conceptos generales biológicos como la *transducción de señales*, *regulación de la expresión genética*, la *comunicación intercelular*. Sen et al. [182] identificaron grupos de proteínas para la levadura de los vectores propios de la matriz Laplaciana (sección 2.1.2.3, capitulo anterior), calculados con *Singular Value Decomposition* **SDV**.

Lewis *et al.* [183] exploraron cuidadosamente la relación entre las comunidades estructurales de las PIN y su función biológica. Las comunidades se detectaron con el enfoque de multiresolución de Reichardt y Bornholt [161] (sección 3.3.5.1). Una comunidad se considera biológicamente homogénea si la similitud funcional entre pares de proteínas de la comunidad (extraída a través de la base de datos *Gene Ontology* [184]) es mayor que la similitud funcional entre todos los pares de proteínas de la red. Lewis *et al.* especificaron también la comparación a pares de interacción de proteínas y no interactuantes. Como resultado, muchas comunidades resultan ser biológicamente homogéneas, sobre todo si no son demasiado pequeñas. Además, algunos atributos topológicos de las comunidades, como el c*oeficiente de agrupamiento* (sección 2.2.3) dentro de la comunidad (es decir, el valor medio de los coeficientes de la agrupación de los nodos de una comunidad, teniendo en cuenta sólo a los vecinos que pertenecen a esta) y la *densidad de enlace* (densidad de aristas internas), son buenos indicadores de la homogeneidad biológica: el primero está fuertemente correlacionado con la homogeneidad biológica, independientemente del tamaño de la comunidad, mientras que para este último la correlación es fuerte para grandes comunidades.

Redes Metabólicas y genéticas.

Las redes metabólicas también han sido ampliamente investigadas. Guimerà y Amaral encontraron módulos en una representación cartografica de la red metabolica de *E. coli* [185] con un algoritmo basado en *Simulated annealing* [145] (sección 3.3.4.3). Una técnica de *descomposición jerárquica* de redes metabólicas se ha derivado de Holme *et al.* [186], mediante el uso de una técnica de *agrupación jerárquica* (sección 3.3.1) inspirada en el algoritmo de Girvan y Newman (sección 3.3.2.1) donde los nodos se eliminan con base en sus valores de *intermediación*. Así, emerge un cuadro de la red metabólica, en el que hay grupos principales centrados en sustancias *hub*, rodeadas por capas externas de sustancias menos conectadas y unos pocos grupos en otras escalas intermedias. En general, los grupos en diferentes escalas parecen ser significativos, por lo que toda la jerarquía deberían tenerse en cuenta.

Las *redes de regulación transcripcional* también muestran modularidad en varios organismos [187, 188, 189] (ver sección 1.1.2 capitulo 1). Mucho del trabajo reportado es para pequeños organismos como *virus*, *procariotas* y *levaduras* pero aún no hay mucho para organismos eucariontes y humano. Algunos trabajos analizan modularidad en redes de regulación genética booleanas [190, 191], pero sin incorporar la idea de *Redes Complejas* y también hay muchos basados en asociar funciones a procesos biologicos mediante el uso de *motifs* [192, 193] que son sub-redes mucho más pequeñas asociadas a circuitos genéticos [194].

Wilkinson y Huberman [195] analizaron una red de genes de *co-ocurrencia* para encontrar grupos de genes relacionados. La red se construye mediante la conexión de pares de genes que se mencionan juntos en el resumen de artículos de la base de datos *Medline* (http://medline.cos.com/). Las comunidades se encontraron con una versión modificada del algoritmo de Girvan y Newman [157] en el que la intermediación de arista se calcula teniendo en cuenta los caminos más cortos de un subconjunto pequeño de todos los pares de nodos [196], para ganar tiempo de computo (sección 3.3.2.2). Como resultado, los genes pertenecientes al mismo modulo resultan estar relacionadas entre sí funcionalmente. La co-ocurrencia de términos también se utiliza para extraer asociaciones entre genes y enfermedades, para averiguar qué genes son relevantes para una enfermedad específica.

Es sobresaliente la cantidad de trabajo publicado para el caso de la levadura (*Saccharomyces cerevisiae*) (sección 1.1.1) [197, 198, 199, 200], donde la busqueda de módulos en *GRN* ha llevado a desarrollar métodos basados en *algoritmos de agrupamiento no supervisados* (sección 3.3.1) en experimentos de expresión de genes. Por ejemplo Bar-Joseph *et al.* [201] propusieron un algoritmo para descubrir módulos que combina información de conjuntos de datos de *ubicación* y *expresión* de genoma en *S. cerevisiae*. Un módulo se define como un *conjunto de genes coexpresados* a los que se une el mismo conjunto de factores de transcripción.

Sin embargo la modularidad en *Redes de Regulación Genética* (sección 1.1.2) también se ha estudiado en otros organismos, por ejemplo, Oliveira *et al.* [202] estudiaron la modularidad en una red regulatoria de baculovirus. Freyre *et al.* [203] aplican el método de *hierarchical clustering* de Rives and Galitski [177] previamente usado por O. Resendis *et al.* [204] y lo comapran con los resultados obtenidos del algoritmo de Girvan y Newman [157] e *Infomap* [156] para develar la organización modular de la red transcripcional de *B. Subtilis*. Sanz *et al.* [205] aplican ideas de optimización de modularidad (sección 3.3.4) detectar comunidades en redes de regulación transcripcional en bacterias y Chauhan *et al.* [206] detectaron la estructura modular en red de regulación del factor Sigma de *Mycobacterium tuberculosis* usando el método de Barber et. al. [207] para comunidades en redes bipartitas. Asimismo, para el caso de la planta *Arabidopsis thaliana* Mao *et al.* [208] usaron el *Algoritmo de Agrupamiento de Markov* de Van Dongen [209] (sección 3.3.5.4) para encontrar *módulos funcionales*.

Para el caso de redes de regulación genética en humano podemos encontrar uno de los resultados del presente trabajo donde Alcalá-Corona *et al.* identificaron módulos en una red transcripcional del Factor de Transcripción MEF2C obteniendo funciones biológicas estadísticamente significativas asociadas a dichos módulos [210]. La red fue constituida de manera teorice a partir del proyecto FANTOM4 [211] y los módulos fueron detectados usando el algoritmo *Infomap* [156] (secciones 4.2.1 y 5.1.1). En el caso particular del estudio de cáncer podemos encontrar el trabajo de Srivastava et. al. [212] usaron el método de Clauset y Newman [213] de optimización de la modularidad Q (sección 3.3.4) en una red bipartita (sección 2.1.1.1) de genes y proteinas de cáncer de mama, para encontrar módulos y analizar procesos biológicos relacionados a los mismos usando GeneOntology [184]. Shi *et al.* desarrollaron un algoritmo para identificar *cliques máximos* como módulos de coexpresión para el análisis de datos de expresión génica en cáncer de mama [214] (ver sección 1.2.1 capitulo 1). Y finalmente Alcalá-Corona *et al.* encontraron módulos asociados a procesos biológicos en redes de subtipos de cáncer de mama [215].

Relevancia y aplicaciones de la detección de comunidades en redes complejas.

Como queda de manifiesto, las comunidades (o módulos) son una característica topológica inherente las redes y a la complejidad de los fenómenos que representan, identificarlas puede tener aplicaciones concretas así como implicaciones teóricas importantes, que pueden ir más allá de localizar grupos en redes sociales. Por sólo mencionar un ejemplo en redes informáticas, identificar grupos de clientes en la *Web* que tienen intereses similares y son geográficamente cercanos entre si, puede mejorar el rendimiento de los servicios prestados en la misma, dado que cada grupo de clientes podría estar atendido por un servidor espejo dedicado.

Además, analizar toplógicamente una red y su estructura modular, permite una clasificación de los nodos de acuerdo con su posición estructural en su propia comunidad. Por lo tanto, los nodos con una posición central en sus grupos, *i.e.* que comparten una gran cantidad de aristas con sus compañeros de modulo, puede tener una función importante de control y estabilidad en la comunidad; así también los nodos situados en los límites entre los módulos pueden desempeñar un papel de mediación y conducir las relaciones e intercambios entre diferentes comunidades[216]. Asimismo, se puede estudiar la red donde los nodos son las comunidades y las aristas se fijan entre los grupos si hay conexiones entre algunos de sus nodos en la red original, y/o si se superponen los módulos. De esta manera se alcanza una descripción mesoscopica (o de *coarse graining*) de la red original, que revela relaciones entre módulos. Se ha reportado que las redes de comunidades tienen una distribución de grado diferente con respecto a las redes en su totalidad [175].

Otro aspecto importante relacionado con la estructura modular, es la organización jerárquica mostrada por la mayoría de los sistemas en red. Las redes reales generalmente están compuestas por comunidades, incluyendo comunidades más pequeñas, las cuales a su vez incluyen comunidades más pequeñas, etc. La generación y evolución de un sistema organizado en subsistemas interrelacionados estables es mucho más rápida que si el sistema estuviera desestructurado, porque es mucho más fácil de ensamblar las subpartes más pequeñas primero y usarlas como bloques de construcción para conseguir estructuras más grandes, hasta que todo el sistema está ensamblado. De esta manera también es mucho más difícil que errores (por ejemplo, afectaciones por mutaciones en redes transcripcionales) ocurran a lo largo del proceso.

Asimismo, como es el caso del estudio presentado en este trabajo, identificar módulos en redes de regulación genética puede asociar y descubrir funciones subyacentes en el intrincado programa de regulación genética. Antes de presentar nuestra metodología y resultados en el capitulo siguiente, presentaremos a continuación algunos elementos formales y matemáticos de la detección de comunidades en redes complejas. Como veremos en las próximas secciones, la detección de comunidades en redes complejas sigue siendo un problema no cerrado en Ciencias de la Computación [217] y hay una gran variedad de métodos de detección y algoritmos [146, 147, 218, 219, 220]; por lo tanto, la estructura modular es un tema de especial relevancia en el caso de las redes de regulación transcripcional de genes [221], donde los módulos o comunidades corresponden a conjuntos de genes co-regulados [170, 195, 222, 223].

De esta manera, el problema de detección de comunidades ha llamado fuertemente la atención en la literatura abriendo todo un campo nuevo en la *Ciencia de Redes*. El trabajo de varios académicos se ha enfocado en estudiar, por ejemplo, si las comunidades emergen de los *patrones de mezcla* (sección 2.2.5.2) [224] y en encontrar cada vez mejores algoritmos para detectar comunidades [147].

Elementos de detección de Comunidades

El problema de la detección de la comunidades, es intuitivamente claro a primera vista. Sin embargo, los elementos principales del problema, *i.e.* los conceptos de comunidad y partición, no están rigurosamente definidos, y requieren un cierto grado de arbitrariedad y/o de sentido común. En primer lugar, no hay una única forma de traducir de forma precisa la idea intuitiva de comunidad. De aquí, que el término *Detección de Comunidades* en realidad indica varios problemas diferentes. De hecho, dentro del concepto intuitivo de *comunidad* o *módulo* están ocultas algunas ambigüedades y a menudo hay muchas maneras igualmente legítimas de resolverlas y por lo tanto, no es de extrañar que en la literatura no se aterricen algunas definiciones comunes.

Asimismo, uno de los problemas más importantes, es sin duda, decidir ¿cuál es la mejor partición? Si se pudiera, en principio, analizar todas las posibles particiones de una red, se necesitaría entonces una forma sensata de medir su "calidad" para elegir las particiones con la *estructura modular* más fuerte. Incluso puede ocurrir que una red no tenga estructura modular y entonces se debe ser capaz de considerarlo. Así, encontrar un método para comparar particiones, no es una tarea trivial y hay que idear diferentes opciones para encontrar "buenas" particiones en un tiempo razonable, de nuevo un problema muy difícil.

También es importante resaltar que la identificación de módulos estructurales sólo es posible si las redes son escuetas (sparse) *i.e.* si el número m de aristas es del orden o a lo más un orden de magnitud mayor al de la cantidad de nodos de la red n. Si $m \gg n$, la distribución de aristas entre los nodos es demasiado homogénea para que las comunidades tengan sentido. En dado caso, el problema se convierte en algo bastante diferente, cercano a la agrupación de datos, que requiere de conceptos y métodos de una naturaleza diferente[1]. La principal diferencia es que, mientras que las comunidades en redes están relacionadas, explícita o implícitamente, con el concepto de densidad de de arista (adentro *vs.* afuera de la comunidad), en el agrupamiento de datos las comunidades son conjuntos de puntos que están "cerca" uno del otro, respecto a una medida de distancia o similitud, que se define para cada par de puntos.

De esta manera, en esta sección tiene como objetivo dar una exposición ordenada de los conceptos básicos y fundamentales de la detección de comunidades y discutir las preguntas pertinentes.

[1] algo más cercano a la técnicas tradicionales como la *Agrupación Jerarquica no supervisada* (ver sección 3.3.1)

Definiciones formales de Modulo.

El primer problema dentro de la detección de módulos, es definir con precisión qué es un módulo. Aunque no hay una definición formal, en la literatura es ampliamente aceptado que una comunidad es un conjunto de nodos más conectados entre si que con otros grupos de nodos, es decir la densidad de aristas *intra-comunidad* (aristas dentro un mismo módulo), es mucho mayor que la *inter-comunidad* (densidad de aristas entre módulos). Sin embargo, este concepto se puede formalizar en muchos aspectos, por ejemplo, los analistas de redes sociales han elaborado muchas definiciones de subgrupos con diferentes grados de cohesión interna entre los nodos, así también otras definiciones han sido presentadas por científicos de la computación y físicos. En general, las definiciones se pueden clasificar en tres categorías principales: locales, basadas en similitud nodo y globales.

En las ***Definiciones locales*** la atención se centra en los nodos de la sub-red en estudio y en su vecindad inmediata, sin tener en cuenta el resto de la red. Estos conceptos provienen principalmente del análisis de redes sociales y se pueden subdividir en *auto-referentes*, si se considera solo la sub-red, y *comparativas* cuando la cohesión recíproca de los nodos del sub-red se compara con la cohesión de los vecinos externos. En las ***Definiciones basadas en similitud nodo***, los módulos son grupos de nodos que son similares entre sí. Se elige un criterio cuantitativo para evaluar la similitud entre cada par de nodos, conectados o no. El criterio puede ser local o global: por ejemplo se puede estimar la distancia entre un par de nodos. Las similitudes también pueden ser extraídas de los vectores propios de matrices especiales, que normalmente tienen un valor cercano para nodos pertenecientes al mismo módulo o comunidad (ver sección 3.3.3). Las ***Definiciones globales*** suelen comenzar a partir de un ***modelo nulo***, *i.e.* una red que coincide con la original en algunas de sus características topológicas, pero que no muestra estructura modular. Después de esto, las propiedades de enlace de las sub-redes de la red inicial se comparan con las de las sub-redes correspondientes en el modelo nulo. La forma más sencilla de diseñar un modelo nulo es introducir aleatoriedad en la distribución de aristas. Un red aleatoria tipo Erdös-Rényi, por ejemplo, no se espera que tenga estructura modular, dado que cualquier par nodos tienen la misma probabilidad de ser adyacentes y no hay vinculación preferencial que involucre a grupos especiales de nodos. El modelo nulo más popular es el propuesto por Newman y Girvan [225] y consiste en una versión aleatoria del red original, donde las aristas son reconectadas de forma aleatoria, bajo la restricción de que cada nodo mantenga su grado. Este modelo nulo es el concepto básico detrás de la definición de modularidad, una función que evalúa la calidad de las particiones de una red en módulos (ver sección 3.2.3). Aquí un subconjunto de nodos es un módulo si el número de aristas dentro del subconjunto supera el número esperado de aristas internas que el subconjunto tendría en el *modelo nulo*.

Vale la pena señalar que, a pesar de la amplia variedad de definiciones, en muchos

algoritmos de detección de módulos, estás no están definidas en absoluto, sino que son un subproducto del procedimiento. Este es el caso por ejemplo de los algoritmos divisivos descritos en la sección 3.3.2 y de los algoritmos dinámicos de la sección 3.3.5.

Complejidad Computacional del problema.

Dado que muchas veces la redes reales estudiadas son muy grandes, la cuestión de la eficiencia en los algoritmos de detección de módulos es algo esencial. La complejidad computacional de un algoritmo es la estimación de la cantidad de recursos requeridos por el algoritmo para realizar una tarea. Esto involucra tanto el número de pasos de cálculo necesarios como el número de unidades de memoria que deben ser asignadas simultáneamente para ejecutar el cálculo. Tales demandas se expresan generalmente en términos proporcionales al tamaño del sistema en estudio. En el caso de una red, el tamaño se indica generalmente por el número de n nodos y/o el número de m aristas. La complejidad computacional de un algoritmo no siempre se puede calcular, de hecho, a veces esta es una tarea muy difícil, o incluso imposible. En estos casos, sin embargo es importante tener al menos una estimación de la complejidad del peor caso del algoritmo, que es la cantidad de recursos computacionales necesarios para ejecutar el algoritmo en el caso más desfavorable para un tamaño determinado del sistema.

Así entonces, la notación $O(n^{\alpha}m^{\beta})$ indica que el tiempo de cálculo, el cual crece como una potencia tanto del número de nodos como de aristas, con exponentes α y β, respectivamente. Idealmente, se quisiera tener los valores más bajos posibles para los exponentes, que corresponderían a los valores más bajos posibles de demandas computacionales. Ejemplos de la Web, con millones de nodos y miles de millones de aristas, no pueden abordarse por medio de algoritmos cuyo tiempo de ejecución crezca más rápido que el orden de el número de nodos $O(n)$ o el orden del número de aristas $O(m)$.

Medida de *Modularidad* (evaluación de particiones).

En sentido estricto, la partición de un red en comunidades (módulos) es la división de la red en grupos, donde cada nodo es asignado a un solo grupo. La última condición puede no ser tan estricta, como se muestra en la sección 3.2.4.3. Para cualquiera que sea la definición de comunidad, hay normalmente un gran número de posibles particiones y es necesario entonces establecer qué particiones exhiben una estructura modular real. Para ello, es necesario una función de "calidad", es decir, un criterio cuantitativo para evaluar qué tan buena es una partición. Newman y Girvan [225] propusieron una primera medida matemáticamente formal de modularidad en redes, la cual ha cobrado gran relevancia en la literatura.

Dada una ***partición*** de la red en comunidades:

$$Q = \sum_{i}(e_{ii} - a_i^2) = Tr(\mathbb{E}) - ||\mathbb{E}^2|| \tag{3.1}$$

Donde cada entrada e_{ij} de la matriz $\mathbb{E}$ es la fracción de todas las aristas de la red, que une los nodos en la comunidad i a los nodos de la comunidad j, $a_i = \sum_j e_{ij}$ y para una matriz arbitraria $\mathbb{X}$, $||\mathbb{X}|| = \sum_i \sum_j x_{ij}$.

Q mide la fracción de las aristas de la red que conectan nodos dentro de la misma comunidad (aristas *intra-comunidad*) menos el valor esperado de la misma cantidad en una red con las mismas comunidades (*i.e.* la misma *partición*), pero con conexiones aleatorias entre los nodos. Si el número de aristas *intra-comunidad* no es mejor que el *modelo nulo* (aleatorio), se obtiene que $Q = 0$ y por el contrario si el valor se acerca al máximo $Q = 1$, indica que la red tiene una fuerte estructura modular.

Esta medida puede ser reescrita[213] como:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j) \tag{3.2}$$

Donde m es el número total de aristas de la red, k_i es el grado del nodo i, A_{ij} son los elementos de la matriz de adyacencia y si los nodos i y j **están en la misma comunidad** (*i.e.* $C_i = C_j$) entonces la *función-δ_{ij} de Kronecker*, $\delta(C_i, C_j) = 1$ y si no lo están ($C_i \neq C_j$) entonces $\delta(C_i, C_j) = 0$.

Q también se puede escribir[226, 227] como:

$$Q = \sum_{s=1}^{p} \left[\frac{m_s}{M} - \left(\frac{K_s}{2M} \right)^2 \right] \tag{3.3}$$

Donde la suma es sobre los p módulos (o comunidades) de la partición, m_s es el número de aristas dentro del modulo s, M es el número total de aristas en la red y K_s es el grado total de los nodos en el módulo s. De tal manera que el primer término de la suma es la fracción de aristas dentro del módulo s y el segundo término, en cambio, representa la fracción esperada de aristas en ese módulo, si los enlaces fueran colocados aleatoriamente en la red, bajo la única restricción de que la secuencia de grado coincida con el de la red original (*modelo nulo*). La ecuación (3.3) incorpora una definición implícita de módulo: una sub-red es un módulo, si el número de aristas en su interior es mayor que el número esperado en un modelo de *modularidad nula*. Si este es el caso, los nodos de la sub-red están más estrechamente conectados de lo esperado. Básicamente, si cada sumando de la ecuación (3.3) es positivo, la sub-red correspondiente es un módulo. Cuanto mayor sea la diferencia entre las aristas reales y las esperadas, mayor sera la *modularidad* de la sub-red, por lo que se espera que grandes valores positivos de Q

indiquen buenas particiones.

Así también Q se puede escribir la modularidad para redes pesadas [228] como:

$$Q_w = \frac{1}{2W} \sum_{i,j} \left[W_{ij} - \frac{s_i s_j}{2W} \right] \delta(C_i, C_j) \tag{3.4}$$

Donde, para una normalización adecuada, la cantidad de aristas m de la ecuación (3.1) se reemplaza por la suma W de los pesos de todos las aristas. El producto $\frac{s_i s_j}{2W}$ ahora es el peso esperado de la arista ij en el *modelo nulo* de modularidad, que tiene que ser comparado con el peso real W_{ij} de esa arista en la red original.

La *modularidad* Q siempre es menor que 1, y también puede ser negativa. Por ejemplo, la partición en la que cada nodo es una comunidad es siempre negativa y asimismo la modularidad de toda la red, vista como una sola comunidad, es cero, ya que en este caso los dos términos en cada sumando son iguales y opuestos. Esta es una buena característica de la medida, ya que implica que, si no hay particiones con modularidad positiva, la red no tiene estructura de modular. Por el contrario, la existencia de particiones con grandes valores negativos de modularidad puede aludir a la existencia de subgrupos con muy pocas aristas internas y muchas aristas situadas entre ellos (estructura *multipartita*).

La modularidad ha sido empleada como una función de de calidad en muchos métodos, tal es el caso de algunos algoritmos divisivos (sección 3.3.2). Además, la optimización de la modularidad en sí es un método popular para detectar de comunidades (sección 3.3.4). La modularidad también permite evaluar la estabilidad de las particiones [229] y transformar una red en otra más pequeña, preservando su estructura modular [230]. Sin embargo, hay algunas limitaciones en el uso de la medida, las más importantes conciernen al valor de la modularidad para una partición *i.e.* ¿para qué valores se puede decir que hay una clara estructura modular en un red? La pregunta es difícil: si dos redes tienen el mismo tipo de estructura modular pero de diferentes tamaños, la modularidad será más grande para el red más grande, así, los valores de modularidad no pueden compararse para redes diferentes.

Por otra parte, cabría esperar que particiones de redes aleatorias tengan valores cercanos a cero de modularidad, ya que en estás, no se espera estructura modular alguna. Sin embargo se ha demostrado que particiones de redes aleatorias pueden alcanzar valores de modularidad bastante grandes, ya que para casos específicos no es despreciable la probabilidad de que la distribución de aristas sea localmente heterogénea [145]. Asimismo, Fortunato *et. al.* han demostrado que la modularidad aumenta si sub-redes menores a un tamaño característico se unen [226]. Este hecho representa un serio sesgo cuando se buscan comunidades a través de la optimización de modularidad.

Finalmente existen otras medidas de modularidad propuestas como por ejemplo la propuesta por Reichardt y Bornholdt [161] basándose en la medida Q original y en una medida de calidad presentada por ellos mismos en un trabajo anterior [160], Asimismo hay otras formas de buscar la mejor estructura modular en una red compleja como la presentada por Rosvall y Bergstrom en su método *Infomap* [156], al maximizar la descripción codificada de una caminata aleatoria en la red.

Consideraciones especiales de la Estructura Modular.

Existen consideraciones en la estructura modular que pueden llevar a ambigüedades. Por ejemplo, la calidad de una partición puede llegar a ser distinta para redes pesadas y/o dirigidas por lo que los métodos deben tener esta consideración, asimismo al detectar comunidades en redes bipartitas o multipartitas. También, algunas definiciones permiten que los nodos pertenezcan a más de una comunidad, con lo que se tienen comunidades traslapadas y no traslapadas. Otra ambigüedad tiene que ver con el concepto mismo de *estructura modular*, dado que se puede pensar como una sola partición simple de la red o como una jerarquía de particiones en diferentes niveles o (*coarse graining*).

Redes Dirigidas y Pesadas.

En algunas de las redes expuestas en la sección 3.1.1 las aristas tienen (o pueden tener) pesos. Por ejemplo, los aristas de la red de colaboración de la 3.3 panel **b**, podrían ser pesadas por el número de trabajos en coautoría con pares de científicos. Los pesos son valiosa información adicional en una red, y se debe considerar en el análisis, en general los métodos para en redes no pesadas se pueden extender de forma simple al caso pesado, caso contrario al caso de las redes dirigidas. En la *World Wide Web*, por ejemplo, se puede pasar de la página A a la página B haciendo clic en un hipervínculo de A, pero por lo general no se encuentra en un hipervínculo B que regrese a A[1].

Las comunidades de la *Web* pueden ser grupos de páginas que tienen temáticas similares y en general se supone que la existencia de un enlace entre dos páginas implica que ambas tienen contenido relacionado y que dicha relación es independiente de la dirección de hipervínculo, por lo que en algunos estudios realizados de detección de comunidades se ha descuidado la direccionalidad de los hipervínculos y se ha examinado la red como no dirigida. Sin embargo, independientemente de la red, tener en cuenta la direccionalidad de las aristas puede mejorar considerablemente la calidad de la(s) partición(es), así como develar mucha información valiosa acerca del sistema. A pesar de esto, el desarrollo de métodos de detección de comunidades para redes dirigidas es una tarea difícil, ya que estas se caracterizan por matrices asimétricas (de adyacencia, laplaciana, etc.) así por ejemplo, el análisis espectral es mucho más complejo (ver sección

[1]De hecho, muy pocos hipervínculos (menos del 10 %) son recíprocos

3.3.3). Sólo unas pocas técnicas se pueden extender fácilmente a partir del análisis no dirigido al caso dirigido, en otros casos el problema debe ser formulado a partir de cero.

Redes Multipartitas.

Hasta ahora hemos hablado de ejemplos en redes unipartitas. Sin embargo, no es raro encontrar redes reales con diferentes clases de nodos así como aristas que unen nodos de diferentes clases. Un ejemplo es una red de científicos y artículos, donde las aristas unen a científicos con los artículos que han publicado. Para una red multipartita el concepto de comunidad no cambia mucho con respecto al caso de redes unipartitas, ya que permanece relacionado con una gran densidad de aristas entre los miembros del mismo grupo, con la única diferencia de que los elementos de cada grupo pertenecen a diferentes clases de nodos. Las redes multipartitas normalmente se reducen a las proyecciones unipartitas de cada clase de nodo (véase sección 2.1.1.1). Por ejemplo, a partir de la red bipartita de los científicos y los artículos, se puede extraer una red de sólo científicos, que están relacionados por coautoría.

La detección de comunidades en redes multipartitas puede tener aplicaciones interesantes y aunque se pueden adoptar métodos estándar de análisis comunidades, una gran cantidad de información se pierde. Asi, el problema de la detección de comunidades en redes multipartitas no es trivial, y por lo general requiere de metodologías *ad hoc*.

Jerárquica y superposición de módulos.

Los módulos en si mismos, pueden mostrar también una estructura modular interna, *i.e.* pueden contener módulos más pequeños, los que a su vez pueden incluir otros módulos, y así sucesivamente. En este caso se dice que la red es jerárquica (ver figura 3.6). Para una clasificación clara de los nodos y sus roles dentro de un red, es importante encontrar todos los módulos de la red, así como su jerarquía.

Una forma natural para representar la estructura jerárquica de una red es dibujar un dendrograma, como el que se ilustra en la figura 3.9. En este caso, se muestran las particiones de una red con doce nodos. En la parte inferior, cada nodo es su propio módulo. Al moverse hacia arriba, los grupos de nodos se agregan sucesivamente. Las fusiones de comunidades están representadas por líneas horizontales. El nivel más alto representa la red entera como una sola comunidad. Cortar el diagrama horizontalmente a cierta altura, como se muestra en la figura 3.9. (líneas discontinuas), muestra un nivel de organización de los nodos en la red. El diagrama es jerárquico por construcción[1]: cada comunidad perteneciente a un nivel es incluido totalmente en una comunidad en un nivel superior.

[1]En general mediante agrupación jerárquica no supervisada (ver sección 3.3.1).

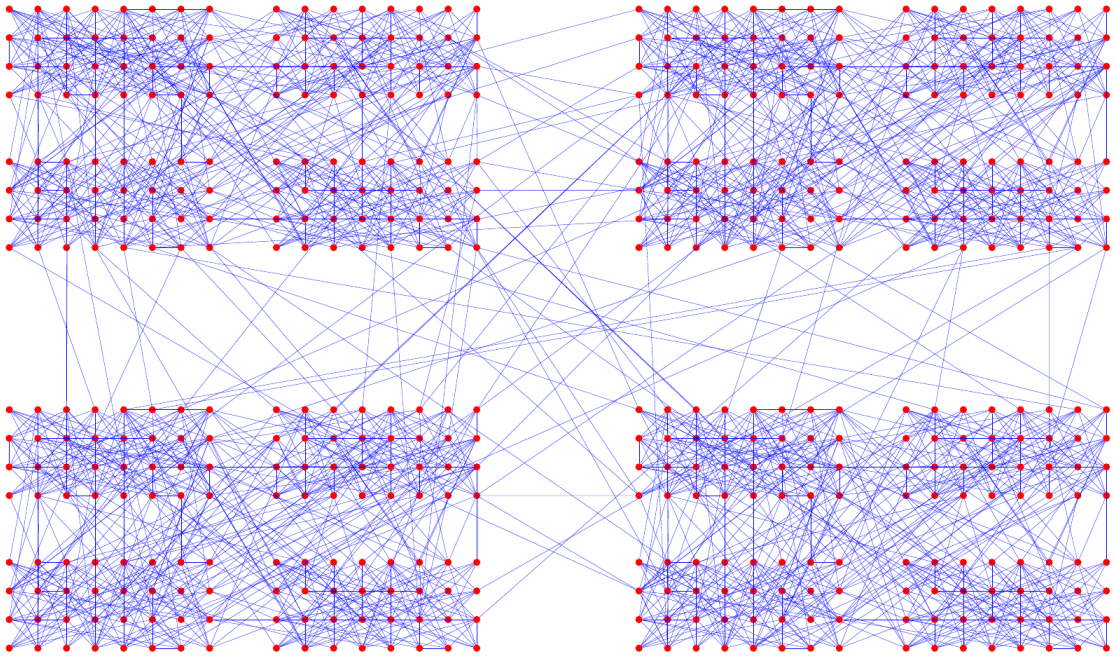


Figura 3.6: **Ejemplo esquemático de una red con estructura modular jerárquica**. La red consta de cuatro módulos principales, cada uno con 4 submódulos. Los 16 submódulos constan de 32 nodos con grado $k = 64$.

En una partición estándar cada nodo se atribuye generalmente a un solo módulo. Sin embargo, los nodos situados en la frontera entre los módulos pueden ser difíciles de asignar a un módulo u otro, basandose en sus conexiones con los otros nodos. En este caso, tiene sentido considerar a tales nodos intermedios como pertenecientes a más de un sólo módulo, con lo que podemos tener comunidades superpuestas o traslapadas (figura 3.7) y utilizar el término *cubierta*, en lugar del de partición, cuya definición estándar prohíbe una membresía múltiple de nodos. Los algoritmos estándar de detección de comunidades asignan cada nodo a un único módulo y al hacerlo, pierden información potencialmente relevante o incluso a menudo discrepan en la clasificación de nodos periféricos de módulos, ya que están obligados a ponerlos en un solo grupo, lo que podría ser incorrecto.

Muchas redes reales se caracterizan por una estructura modular con superposiciones importantes entre diferentes grupos, donde los nodos pueden pertenecer a más de un módulo. Ejemplos clásicos son las redes sociales donde las personas suelen pertenecer a más de una comunidad, como compañeros de trabajo, familia, etc. Tener en cuenta las superposiciones es también una manera de explotar mejor la información que se puede derivar de la topología, ya que los nodos pertenecientes a varios módulos pueden

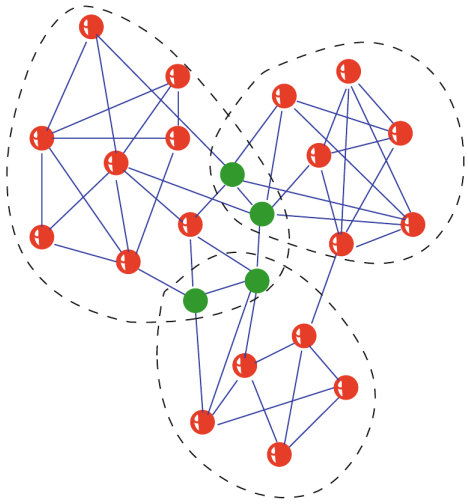


Figura 3.7: **Comunidades superpuestas (o traslapadas).** En la partición resaltada por los contornos discontinuos, algunos nodos se comparten entre más de una comunidad.

desempeñar un papel importante de intermediación entre las diferentes comunidades de la red e incluso se podría estimar el grado de participación de un nodo en diferentes módulos. Sin embargo, tomar en cuenta las superposición de módulos introduce una variable adicional, la *membresía* de los nodos en diferentes comunidades, lo que aumenta enormemente la cantidad de cubiertas posibles con respecto a las particiones estándar. Por lo tanto, la búsqueda de comunidades superpuestas suele ser computacionalmente mucho más exigente que la detección de particiones estándar. El problema es tan difícil que muy pocos algoritmos consideran la posibilidad de tener comunidades superpuestas.

Métodos de detección de comunidades.

El problema de particionar una red, consiste en dividir los nodos de la misma en g grupos de tamaño predefinido, tal que el número de aristas situadas entre los grupos sea mínimo. La figura 3.8 presenta la solución del problema para una gráfica con catorce nodos y $g = 2$ grupos de igual tamaño. La mayoría de las variantes del problema de partición de redes son problemas *NP-completos*, por lo que es poco probable que la solución se pueda calcular en un tiempo que crezca como potencia del tamaño n de la red. No obstante, existen varios algoritmos que pueden hacer un buen trabajo, aunque sus soluciones no necesariamente son las mejores.

Dado que el problema de detectar módulos o comunidades en redes complejas es

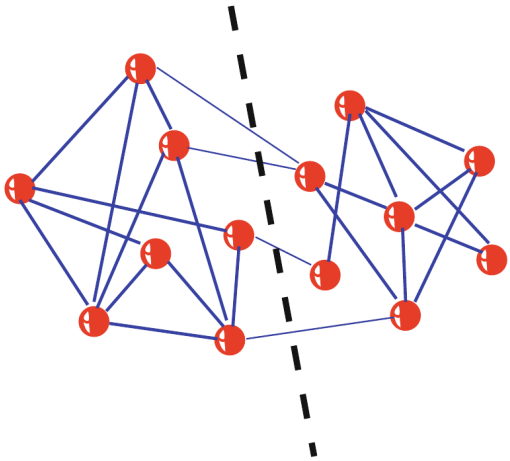


Figura 3.8: **Particionamiento de red.** El corte muestra la partición en dos grupos de igual tamaño.

un problema no cerrado, existen muchísimas técnicas y métodos para poder abordar el problema[231]. Exponer todos o lo mayoría de ellos merecería un libro completo en si, sin embargo existen excelentes revisiones al respecto de Santo Fortunato, una primera de 2010 [146] así como una mas reciente de 2016 [147]. A continuación expondremos solamente algunos de los algoritmos más destacados de algunas de las familias de métodos existentes, de las muchas citadas en la literatura, este recuento no pretende ser exhaustivo por lo que si se requiere más detalle de los métodos o técnicas se sugiere ir a las fuentes originales o bien para tener una mejor panorama ir a los artículos de revisión ya mencionados [146, 147, 231].

La exposición de métodos detallada en está sección está fuertemente basada en las revisiones de Santo Fortunato *et al.* sobre el tema [146, 147], para mayor detalle se sugiere consularlas.

Técnicas Tradicionales: Agrupación jerárquica y k-means clustering.

Comencemos con dos de las técnicas más tradicionales utilizadas para realizar análisis de agrupamiento, derivadas del estudio en redes sociales: el *agrupamiento jerárquico* y el *k-means clustering*.

Agrupación jerárquica. El punto de partida de la *agrupación jerárquica* es la definición de una medida de similitud entre nodos. Después de que una medida es elegida, se calcula la similitud para cada par de nodos, sin importar si están conectados o no. Al final de este proceso, se obtiene una nueva matriz $\mathbb{X}$ de $n \times n$, llamada *matriz de similitud*. Inicialmente, hay n grupos, cada una conteniendo a cada uno de los nodos.

En cada paso, los dos grupos más similares se fusionan; el procedimiento continúa hasta que todos los nodos están en el mismo grupo. Lo anterior se ilustra mejor por medio de dendrogramas, como el de la figura 3.9. Hay diferentes maneras de definir la similitud (o similaridad) entre grupos fuera de la matriz $\mathbb{X}$. Una posibilidad es definir una distancia entre nodos, como:

$$X_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \tag{3.5}$$

Esta es una medida de disimilitud, basada en el concepto de equivalencia estructural. Dos nodos son estructuralmente equivalentes si tienen los mismos vecinos, incluso si ellos mismos no son adyacentes. Si i y j son estructuralmente equivalentes, $X_{ij} = 0$. Los nodos con alto grado y vecinos diferentes se consideran "lejos" unos de otros.

Hay que señalar que la agrupación jerárquica no regresa una sola partición, sino un conjunto de particiones. Sin embargo, tiene la ventaja de que no requiere un conocimiento preliminar sobre el número y tamaño de los grupos (comunidades). Sin embargo, no proporciona una forma de discriminar entre las distintas particiones obtenidas por el procedimiento, y elegir aquel o aquellas que mejor representan la estructura modular de la red. Además, los resultados del método dependen de la medida de similitud específica adoptada. Por último, no se clasifican correctamente todos los nodos de una comunidad, y en muchos casos algunos nodos se pierden, incluso si tienen un papel central en sus grupos [231].

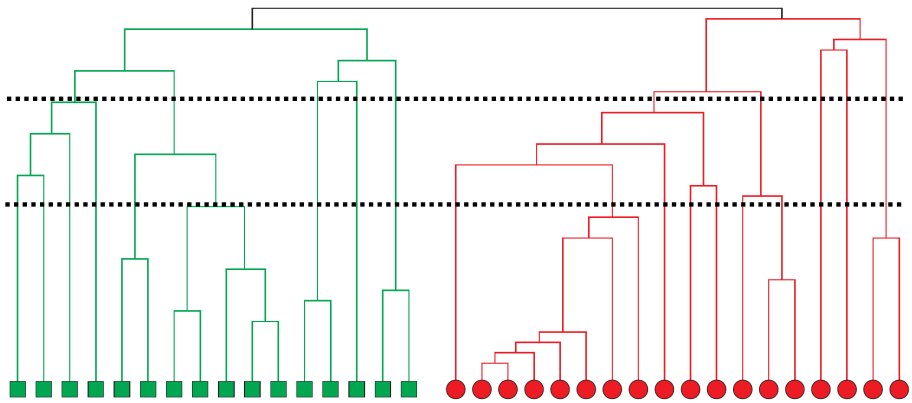


Figura 3.9: **Dendrograma (o árbol jerárquico).** Este dendograma corresponde a la partición de la red del *club de Karate de Zachary* (sección 3.4.1.1) mediante el algoritmo de Girvan-Newman (sección 3.3.2.1). Los cortes horizontales corresponden a las particiones del gráfico en las comunidades.

k-means clustering. Otra técnica de agrupación popular en el estudio de redes sociales es *k-means clustering* [232]. En este caso, el número de grupos está preasignado, por ejemplo k y los nodos de la red se insertan en una espacio métrico, de modo que cada nodo es un punto y una medida de distancia se define entre los pares de puntos en el espacio. La distancia es una medida de disimilitud entre nodos. El objetivo del algoritmo es identificar los k puntos en este espacio, o *centroides*, de modo que cada nodo este asociado a un *centroide* y la suma de las distancias de todos los nodos de sus respectivos *centroides* sea mínima. Para lograr esto, se parte de una distribución inicial de centroides tal que sean lo más lejanos posible entre sí. En la primera iteración, cada nodo está asignado al centroide más cercano. A continuación, se estiman los centros de masa de los k grupos y se convierten en un nuevo conjunto de centroides, lo que permite una nueva clasificación de los nodos, y así sucesivamente. Después de que un número suficiente de iteraciones, las posiciones de los centroides son estables, y los grupos (comunidades) no cambian más. La solución encontrada no es necesariamente óptima, ya que depende fuertemente de la elección inicial de los centroides. El resultado puede ser mejorado mediante la realización de más ejecuciones a partir de diferentes condiciones iniciales. La limitación del *k-means clustering* es que el número de grupos debe ser especificado al principio, pues el método no es capaz de obtenerlos. Además, la incrustación en un espacio métrico puede ser natural para algunos redes, pero bastante artificial para otras.

Es claro que los enfoques tradicionales para obtener particiones de redes tienen serias limitaciones. El problema más importante es la necesidad de proporcionar a los algoritmos la información que a uno le gustaría obtener de los algoritmos mismos, como el número de grupos y su tamaño. Aun cuando estos factores no son necesarios como en el *agrupamiento jerárquico*, existe la cuestión de estimar de la buena calidad de las particiones, de modo que se pueda escoger la mejor. Por estas razones, ha habido un gran esfuerzo en los últimos años para diseñar algoritmos capaces de extraer una información completa sobre la estructura modular de las redes. Estos métodos más recientes se pueden agrupar en diferentes categorías.

Algoritmos Divisivos

Una forma sencilla de identificar comunidades en una red es detectar las aristas que conectan los nodos de las diferentes comunidades y eliminarlos, de manera que los grupos se desconecten el uno del otro. Esta es la filosofía de los algoritmos divisivos, el punto crucial es encontrar una propiedad de las *aristas inter-comunidades* que pueda permitir su identificación. Cualquier método divisivo ofrece muchas particiones, que son por construcción jerárquicas, de manera que pueden representarse mediante dendrogramas (figura 3.9).

El algoritmo de Girvan-Newman.

El algoritmo divisivo más popular es el propuesto por Girvan y Newman [157]. El método también es de importancia histórica, ya que marcó el comienzo de una nueva era en el campo de la detección de comunidades. Aquí las aristas se seleccionan de acuerdo a los valores de las medidas de *intermediación de arista*, estimando la importancia de las aristas de acuerdo con alguna propiedad o proceso que se ejecute en la red.

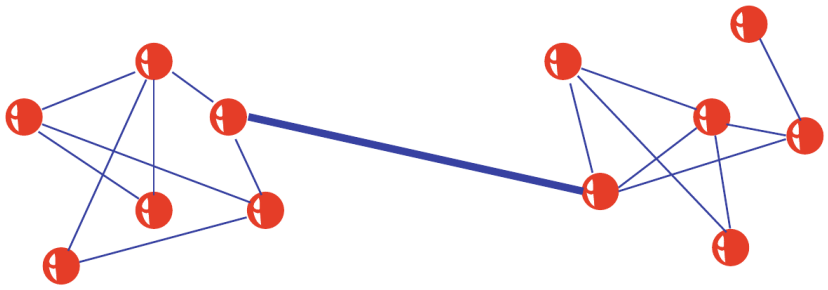


Figura 3.10: **Intermediación de arista (edge betweenness).** La intermediación es más alta para aristas que conectan comunidades. La arista gruesa en el centro tiene una intermediación mucho más alta que todas las demás aristas, porque todos los caminos más cortos que conectan los nodos de las dos comunidades lo atraviesan.

.

La *intermediación de arista de camino corto* [157] es el número de caminos más cortos entre todos los pares de nodos que se pasan por una arista intermediadora y es una extensión a aristas del concepto de intermediación de nodo (ver capitulo anterior, sección 2.2.5.1). Es intuitivo que las aristas intercomunidades tengan un valor grande de intermediación, dado que muchos caminos cortos que conectan nodos de diferentes comunidades pasarán a través de ellas (figura 3.10). Los pasos del algoritmo son:

1. Cálculo de la intermediación para todas las aristas;
2. Eliminación de la arista con mayor centralidad;
3. Nuevo cálculo de la intermediación en la red en curso;
4. Iteración del ciclo desde el paso 2.

Girvan y Newman han considerado dos definiciones alternativas de intermediación: la *intermediación de corriente* e intermediación de *caminata aleatoria* [225]. La *intermediación de corriente* se define considerando a la red como un circuito de resistencias

eléctricas, con aristas que tienen un valor de resistencia. Si un voltaje se aplica entre dos nodos, cada arista transporta una cierta cantidad de corriente, que puede ser calculada mediante la resolución de ecuaciones de Kirchoff. El cálculo de la intermediación de corriente requiere la inversión de una matriz de $n \times n$ (una vez), seguido por la obtención y promedio de la corriente para todos los pares de nodos. Cada una de estas dos tareas toma un tiempo $O(n^3)$ para una matriz *sparse* (o escueta) *i.e.* poco densa. La *intermediación de caminata aleatoria* de una arista, dicta con qué frecuencia un caminante aleatorio sobre la red cruza por la misma. Se calcula la probabilidad de que aristas fueron cruzadas por un caminante, y se promedia sobre todas las opciones posibles para todos los pares de nodos en la red. El cálculo completo requiere un tiempo $O(n^3)$ en una red *sparse* (escueta). Es posible demostrar que esta medida es equivalente a la intermediación de corriente. El cálculo de la *intermediación de camino corto* se puede calcular en un tiempo que escala como $O(mn)$ por lo que es es mucho más rápido que la de flujo de corriente o la de caminata aleatoria. Además, en aplicaciones prácticas del algoritmo Girvan-Newman con intermediación de arista muestra mejores resultados que la adopción de las otras medidas. Estudios numéricos muestran que el re-cálculo de centralidad del paso 3 del algoritmo de Girvan-Newman es esencial para detectar comunidades significativas.

En general, el algoritmo es bastante lento y es aplicable a redes con un máximo de $n \sim 10000$ nodos. En la versión original del algoritmo [157], los autores tuvieron que lidiar con la jerarquía completa de las particiones, ya que no tenía ningún procedimiento para decir qué partición es la mejor.

Otros métodos divisivos.

Otro algoritmo divisivo citado en la literatura [146, 231] es el **algoritmo de Tyler, Wilkinson y Huberman** [196] quienes propusieron un modificación del algoritmo de Girvan-Newman, para mejorar la velocidad del cálculo. La modificación consiste en el cálculo de la contribución a la *intermediación de arista* de solamente de un número limitado de pares de nodos, elegidos al azar, lo que deriva en una especie estimación de *Monte Carlo*. El método se ha aplicado a una red de personas a través de correo electrónico [196] y redes de *co-ocurrencias* de genes [195] (ver sección 3.1.2.2).

Otro más es el **algoritmo de Radicchi *et al.*** [233] que se basa en la idea intuitiva de que dada la alta densidad de aristas dentro de las comunidades, es fácil encontrar ciclos en ellas, *i.e.* rutas cerradas que no se intersectan. Por el contrario, las aristas situadas entre las comunidades difícilmente serán parte de ciclos cortos. Así los autores proponen una medida, el *coeficiente de agrupamiento de arista*, de tal manera que valores bajos de dicha medida probablemente correspondan a aristas entre comunidades. El *coeficiente de agrupamiento de arista* generaliza a las aristas la noción de *coeficiente de agrupamiento* presentado por Watts y Strogatz [115] para nodos (ver capitulo anterior, sección 2.2.3). El coeficiente de agrupamiento de arista es el número de ciclos de longi-

tud g que incluyen a la arista, dividido por el número de ciclos posibles. Por lo general, se consideran ciclos de longitud $g = 3$ o 4. En cada iteración, la arista con el coeficiente de agrupamiento más pequeño se elimina y la medida se vuelve a calcular de nuevo, y así sucesivamente. Puesto que el coeficiente de agrupamiento de arista es una medida local, que involucra a lo más una vecindad ampliada de la de arista, se puede calcular muy rápidamente. Sin embargo, el método puede dar resultados pobres cuando la red tiene pocos ciclos.

Por último tenemos el **Algoritmo de Fortunato, Latora y Marchiori** [234], que se basa en una medida alternativa de centralidad de arista es la *centralidad de información*. Ésta a su vez esta basada en el concepto de *eficiencia* (ver capitulo anterior, sección 2.2.5.1), que estima la facilidad con la que la información viaja en una red de acuerdo con la longitud de los caminos más cortos entre nodos. La centralidad de información en una arista es la variación en la *eficiencia de la red* si la arista es removida. Así, las aristas se eliminan de acuerdo a la disminución de los valores de centralidad de información. El método es análogo al de Girvan y Newman (sección 3.3.2.1), pero más lento.

Hay que señalar que los algoritmos divisivos no son los mejores ni en eficiencia ni en tiempo de ejecución comparados con algoritmos más recientes. Sin embargo es importare remarcar su importancia histórica ya que fueron las primeras propuestas de algoritmos de identificación de comunidades en redes complejas.

Algoritmos Espectrales

Otra familia de técnicas para encontrar particiones en redes se basan en las propiedades espectrales de las matrices que las representan (adyacencia, normalizada, etc.), en particular de la *Matriz Laplaciana*. Como se discutió en la sección 2.1.2.3 del capitulo anterior, la matriz Laplaciana $\mathbb{L}$ (o simplemente Laplaciano) de una red se obtiene de la matriz adyacencia $\mathbb{A}$, colocando en la diagonal los grados de los nodos y cambiando los signos de los otros elementos. El Laplaciano tiene todos los valores propios no negativos y al menos un valor propio cero, ya que la suma de los elementos de cada fila y columna de la matriz es cero. Así, si una red está dividida en g componentes conectados, el Laplaciano tendría g vectores propios degenerados con el valor propio cero y se puede escribir en una forma diagonal por bloques, *i.e.* los nodos pueden ser ordenados de tal manera que el Laplaciano mostrará g bloques cuadrados a lo largo de la diagonal, con entradas diferentes de cero, mientras que todos los demás elementos desaparecen. Cada bloque es el Laplaciano de la sub-red correspondiente, por lo que tiene el vector propio trivial $(1, 1, 1, ..., 1, 1)$. De esta manera, a partir de los vectores propios degenerados se pueden identificar los componentes conectados de la red.

Ahora bien, si la red es conexa, pero consiste en g sub-redes que están débilmente vinculadas entre si, el espectro tendrá un valor propio cero y $g - 1$ valores propios que

son cercanos a cero. Si los grupos son dos, el segundo valor propio más bajo será cercano a cero y el correspondiente vector propio, también llamado *vector de Fiedler* [109, 110] (ver sección 2.1.2.3), se puede utilizar para identificar ambos dos grupos.

Bisección Espectral (spectral bisection method.)

Una de las primeras técnicas que se basan en esta idea es el popular método de *bisección espectral* (*spectral bisection method*), el cual logra particionar la red en dos grupos como se muestra a continuación. Cada partición de una red con n nodos en dos grupos puede ser representada por un vector índice $\vec{s}$, cuya entrada s_i es 1 si el nodo i está en un grupo y -1 si se encuentra en el otro grupo. Así, el tamaño de corte R de la partición del red en dos grupos se puede escribir como:

$$R = \frac{1}{4}\,\vec{s}^{T}\mathbb{L}\vec{s} \tag{3.6}$$

donde $\mathbb{L}$ es la matriz Laplaciana y $\vec{s}^{T}$ el transpuesto del vector $\vec{s}$. Y $\vec{s}$ también se puede escribir como $\vec{s} = \sum_i a_i\vec{v}_i$, donde $i = 1, \ldots, n$ son los eigenvectores del Laplaciano. Si $\vec{s}$ está normalizado apropiadamente entonces:

$$R = \sum_i a_i^2\lambda_i \tag{3.7}$$

donde λ_i es el valor propio Laplaciano correspondiente al vector propio $\vec{v}_i$. Vale la pena señalar que la suma contiene a lo más $n-1$ términos, ya que el Laplaciano tiene al menos un valor propio cero. Este procedimiento proporciona dos particiones: la mejor solución es aquella que da el tamaño corte más pequeño y posteriormente para encontrar las particiones en más de dos módulos se realiza un biseccionamiento iterativo. Sin embargo, utilzar este *bisectionamiento iterativo* para dividir la red en más piezas no es un procedimiento fiable. Este método en general es incapaz de predecir el número y tamaño de los módulos, éstos en su lugar, deben ser proporcionados al procedimiento. Por lo que a continuación, se muestran algunos algoritmos más recientes que son más poderosos.

El algoritmo de Donetti y Muñoz.

Un método elegante basado en los vectores propios de la matriz Laplaciana ha sido propuesto por Donetti y Muñoz [235]. La idea es simple: los valores de las entradas de los vectores propios son cercanos para nodos en la misma comunidad, por lo que se pueden utilizar como coordenadas para representar a los nodos como puntos en un espacio métrico. Por lo tanto, si se usan M vectores propios, se puede insertar los nodos en un espacio M-dimensional. Las comunidades aparecen como grupos de puntos bien separados unos de otros, como se ilustra en la figura. 3.11. La separación es más visible, cuanto mayor sea el número de dimensiones o vectores propios. Los puntos del espacio

se agrupan en comunidades por agrupación jerárquica (véase sección 3.3.1). La última partición es la que tiene mayor modularidad. Como medida de similitud entre nodos, Donetti y Muñoz utilizaron tanto la distancia euclidiana como la distancia angular. El algoritmo se ejecuta hasta completarse en un tiempo $O(n^3)$, lo cual no es muy rápido. Además, el número M de vectores propios que son necesarios para tener una separación limpia en comunidades no se conoce a priori.

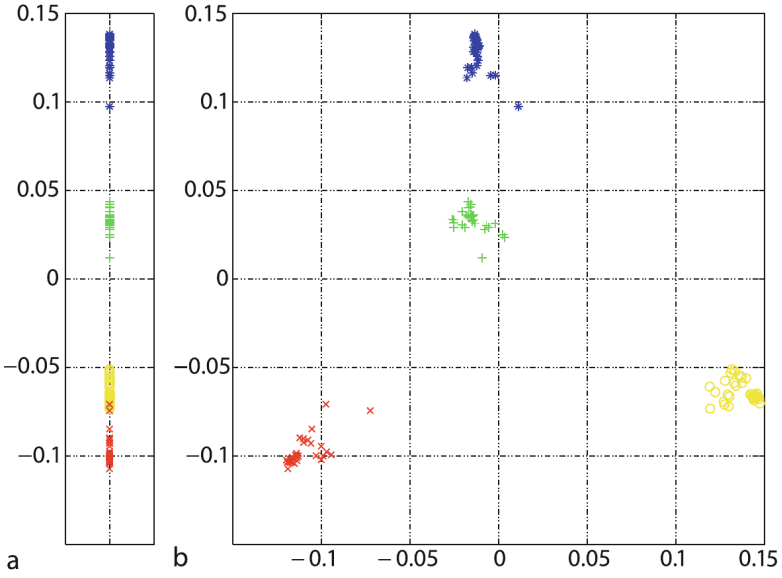


Figura 3.11: **Algoritmo espectral de Donetti y Muñoz.** El nodo i está representado por la i-ésima entrada de los vectores propios de la matriz Laplaciana. En este ejemplo, la red tiene una partición en cuatro módulos, indicados por diferentes símbolos (y colores). Las comunidades están mejor separadas en dos dimensiones (derecha) que en una (izquierda).

Algoritmo de Capocci et al. De manera similar a Donetti y Muñoz, el *el algoritmo de Capocci et al.* [236] utiliza las entradas de los vectores propios para identificar comunidades. En este caso los vectores propios son los de la *matriz normalizada*, que se deriva de la matriz de adyacencia dividiendo cada fila por la suma de sus elementos. Se construye una matriz de similitud mediante el cálculo de la correlación entre las entradas del vector propio. Así la similitud entre los nodos i y j es el *coeficiente de correlación de Pearson* de las entradas de los vectores propios correspondientes. El método se puede extender a redes dirigidas y es útil para estimar similitudes entre nodos, sin embargo, no proporciona una partición bien definida de la red.

Redención de los métodos espectrales.

La *agrupación espectral*, sin embargo, no siempre es confiable. Cuando la red es muy poco densa (escasa o *sparse*), la separación entre los valores propios de los eigenvectores relacionados con la comunidad y la resto no es clara. Los vectores propios correspondientes a valores propios fuera del grueso pueden correlacionarse con nodos de alto grado (*hubs*), en lugar de con la *estructura modular*. Del mismo modo, los eigenvectores relacionados con la comunidad pueden asociarse con valores propios que terminan dentro de la mayoría. En estas situaciones, la selección de vectores propios en función de si sus valores propios asociados están dentro o fuera del grueso produce un conjunto heterogéneo, que contiene información tanto de las comunidades como de otras características. El uso de esos vectores propios para el procedimiento de *agrupamiento espectral* hace que la detección de la comunidad sea más difícil y a veces imposible. Desafortunadamente, muchas de las redes encontradas en estudios reales son muy *sparse* (escasas o escuetas) y pueden conducir a este tipo de problemas.

Para abordar este problema Krazcala *et. al.* [130] introdujeron la ***matriz no de retroceso***, propuesta primeramente por K. Hashimoto [129] (*non-backtracking matrix*) $\mathbb{B}$ donde cada una de las $2m$ aristas dirigidas de recibe una etiqueta de la forma $i \to j$ que indica el par de nodos que conecta y la dirección en la que los conecta:

$$B_{i\to j,k\to l} = \delta_{il}(1 - \delta_{jk}) \tag{3.8}$$

donde δ_{ij} es el *delta Kronecker*. En otras palabras, todos los elementos son cero a menos que la arista $i \to j$ apunte al mismo nodo que la apunta la arista $k \to l$, y las aristas $i \to j$ y $k \to l$ no apunten en direcciones opuestas entre el mismo par de nodos. Los valores propios asociados de los eigenvectores de $\mathbb{B}$ relacionados a una comunidad están separados del grueso hasta el límite de detectabilidad teórica [226], por lo que los métodos espectrales que utilizan los vectores propios superiores de $\mathbb{B}$ que son capaces de encontrar comunidades mientras sean detectables.

Este nuevo enfoque de la teoría espectral se puede encontrar en algoritmos espectrales más recientes, como el propuesto por Krazcala *et. al.* [130] donde, presentan una clase de algoritmo basado en una caminata sin retroceso en las aristas dirigidas de la red. El espectro del operador $\mathbb{B}$ se comporta mucho mejor, manteniendo una fuerte separación entre los valores propios masivos y los valores propios relevantes para la estructura modular, incluso en el caso de redes escasas (*sparse*). Su algoritmo es óptimo para redes generadas por **modelos de bloques estocásticos** (*Stochastic Block Models*, ver sección 3.3.6.2), detectando comunidades dentro del límite teórico. En la misma dirección M.E.J. Newman [237] propone una variante de la matriz de no retroceso $\mathbb{B}$ y estudia el comportamiento de esta matriz para redes artificiales y reales, encontrando que tiene propiedades deseables, especialmente en el caso común de redes con distribuciones de grados amplios, para las cuales parece tener un espectro y vectores propios mejor comportados que la matriz original de no retroceso $\mathbb{B}$. Finalmente Singh y Humphries [238]

presentan los operadores de retroceso **reluctantes** (o "reacios") *reluctant backtracking operators* con los que muestran que se puede detectar comunidades en ciertas redes dispersas tipo árbol (ver capitulo anterior, sección 2.1.1.1) y se desempeñan bien en redes generadas por modelos de bloques estocásticos y redes reales.

Algoritmos basados optimización de Modularidad.

Si la modularidad Q de Newman y Girvan [225] (ver sección 3.2.3) es un buen indicador de la calidad de particiones, entonce la partición correspondiente a su valor máximo para una red determinada debe ser la mejor, o al menos una muy buena. Esta es la motivación principal de la *maximización de modularidad*, quizás la clase de métodos más popular para la detección de comunidades en redes. Una optimización exhaustiva de Q es imposible, debido a la enorme cantidad de formas en las que es posible particionar una red, incluso cuando ésta es pequeña. Además, el máximo real está fuera de alcance, ya que se ha demostrado que la optimización de modularidad es un problema NP-completo [239], por lo que es probable que sea imposible encontrar una solución en un tiempo de crecimiento exponencial con el tamaño de la red. Sin embargo, actualmente hay varios algoritmos capaces de encontrar aproximaciones bastante buenas de modularidad máxima en un tiempo razonable.

El algoritmo de Clauset-Newman.

El primer algoritmo diseñado para maximizar la modularidad es un método aglomerativo (*greedy*) de Clauset y Newman [213], que es un refinamiento posterior del algoritmo original de Girvan y Newman [157]. En este método, se selecciona la partición con el mayor valor de modularidad. Los grupos de nodos son sucesivamente unidos para formar comunidades más grandes de tal manera que la modularidad aumente después de la fusión. Se parte de n grupos, cada uno con un solo nodo. Inicialmente las aristas no están presentes, se agregan una a una durante el procedimiento. Sin embargo, la modularidad siempre se calcula a partir de la topología de toda la red, ya que se desea encontrar sus particiones. Añadir una primera arista al conjunto de nodos desconectados reduce el número de grupos de $n-1$, por lo que ofrece una nueva partición de la red. La arista se elige de modo que esta partición proporcione el máximo aumento de modularidad con respecto a la configuración anterior. Todas las otras aristas se añaden basadonse en el mismo principio. Si la inserción de una de arista no cambia la partición, *i.e.* los grupos son los mismos, la modularidad se mantiene igual. El número de particiones encontradas durante el procedimiento es n, cada uno con un número diferente de agrupaciones, de n a 1. El valor más alto de la modularidad en este subconjunto de particiones es la aproximación a la máxima modularidad dada por el algoritmo. La actualización del valor modularidad en cada iteración se puede realizar en un tiempo $O(n+m)$, por lo que el algoritmo se ejecuta hasta completarse en un tiempo $O((m+n)n)$, u $O(n^2)$ en una red no densa (escueta), lo que es rápido, sin embargo la aproximación que encuentra no es tan buena, en comparación con otras técnicas.

El método de Louvain (Blondel *et al.*)

Sin duda alguna, el método más popular basado en la optimización de la *Modularidad* Q es el *método de Louvain*[1] [162]. Se trata de un enfoque *greedy* diferente para el caso de redes pesadas. Inicialmente, todos los nodos de la red se colocan en diferentes comunidades. El primer paso consiste en un barrido secuencial sobre todos los nodos, dado un nodo i, se calcula la ganancia en la modularidad pesada ecuación (3.4) que proviene de poner i en la comunidad de su vecino y seleccionar la comunidad del vecino que produce el mayor incremento de Q_w, siempre que sea positivo. Al final del barrido, se obtiene la primera partición de nivel. En el segundo paso, las comunidades son reemplazadas por super-nodos, y dos super-nodos están conectadas si hay al menos una arista entre los nodos de las comunidades correspondientes. En este caso, el peso de la arsita entre los super-nodos es la suma de los pesos de las aristas entre las comunidades representadas en el nivel inferior. Luego se repiten los dos pasos del algoritmo, obteniéndose nuevos niveles jerárquicos y super-redes (figura 3.12).

La modularidad siempre se calcula a partir de la topología red inicial y operar en super-redes permite considerar las variaciones de modularidad para las particiones de la red original después de la fusión y/o división de grupos de nodos. Por lo tanto, en alguna iteración, la modularidad no puede aumentar más y el algoritmo se detiene. La técnica está más limitada por las demandas de almacenamiento que por el tiempo computacional. Este último crece como $O(m)$, por lo que el algoritmo es extremadamente rápido y redes con hasta 10^9 aristas pueden analizarse en un tiempo razonable. Los máximos de modularidad encontrados por el método son mejores que los encontrados con la técnica de Clauset *et al.* [213].

Templado simulado (Simulated annealing).

El *templado simulado* [240] es un procedimiento probabilístico de optimización global que se utiliza en diferentes campos y problemas. Consiste en realizar una exploración del espacio de estados posibles, buscando el óptimo global de una función F, por ejemplo su máximo. El *templado simulado* fue empleado por primera vez para la optimización de modularidad por R. Guimerá *et al.* [145]. Su implementación estándar combina dos tipos de "movimientos": *locales*, donde un nodo tomado al azar se traslada de un grupo a otro, y *movimientos globales* consistentes en fusiones y divisiones de las comunidades. El método potencialmente puede llegar muy cerca de la máxima modularidad verdadera, pero es lento. Por lo tanto, se puede utilizar para redes pequeñas, de hasta unos 10^4 nodos. Las aplicaciones incluyen estudios en paisajes de energía potencial [229] y en redes metabólicas [185] (véase sección 3.1.2.2).

[1] en referencia a la **Universidad de Louvain** Francia, donde fue desarrollado por Blondel *et. al.*

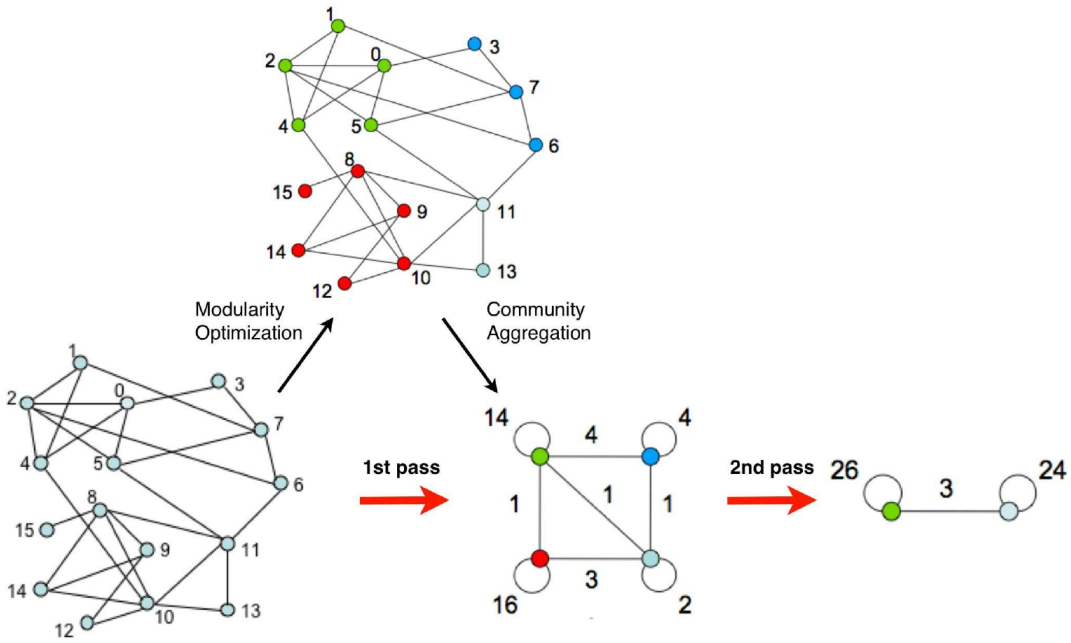


Figura 3.12: **Método de Louvain (Blondel *et al.*).** El diagrama muestra dos iteraciones del método de optimización jerárquica de la modularidad, comenzando por la red de la izquierda. Cada iteración consiste en un paso, en el que cada nodo se asigna a la comunidad (local) que produce el mayor aumento de modularidad, seguido de una transformación sucesiva de los módulos en nodos de una red (pesada) más pequeña, representando el siguiente nivel jerárquico superior.

Optimización espectral (Spectral optimization).

La modularidad se puede optimizar utilizando los valores y vectores propios de una matriz especial llamada *matriz de modularidad* $\mathbb{B}$, cuyos elementos son:

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{3.9}$$

donde la notación es la misma utilizada en la ecuación (3.2). Este método propuesto por Newman [164, 165] es análogo a la *bisección espectral* (sección 3.3.3.1). La diferencia es que aquí la *matriz Laplaciana* se sustituye por la matriz de modularidad $\mathbb{B}$. Entre Q y $\mathbb{B}$ hay la misma relación que entre R y $\mathbb{L}$ en la ecuación. (3.6), por lo que la modularidad se puede escribir como una suma ponderada de los valores propios de $\mathbb{B}$, al igual que en la ecuación (3.7). Aquí se tiene que buscar el vector propio de $\mathbb{B}$ con el valor propio mas alto u_1, y agrupar los nodos de acuerdo a los signos de las entradas de

$\vec{u}_1$. El procedimiento se repite para cada uno de los grupos por separado, y el número de comunidades aumenta tanto como lo hace la modularidad. El algoritmo se ejecuta normalmente en un tiempo $O(n^2 log(n))$ para una red *sparse*.

La ventaja sobre la bisección espectral es que no es necesario especificar el tamaño de los dos grupos, ya que se determinan tomando la partición con mayor modularidad. El inconveniente es similar al utilizado para bisección espectral, *i.e.* el algoritmo da los mejores resultados para bisecciones, mientras que es menos preciso cuando el número de comunidades es mayor a dos. Ésta situación podría mejorarse mediante el uso de los vectores propios con otros valores propios positivos de la matriz de modularidad. Además, los vectores propios con los valores propios más negativos son importantes para detectar una posible estructura multipartita de la red, ya que dan la contribución más relevante para el mínimo de la modularidad. En un trabajo más reciente Zhang y Newman [241] han atacado estos problemas presentando un algoritmo espectral que puede dividir directamente una red en cualquier cantidad de comunidades. El algoritmo utiliza un mapeo de la maximización de la modularidad a un problema de partición de vectores, combinado con una heurística rápida para la partición de vectores.

Límites de la optimización de Modularidad.

Hay que tener, algunas consideraciones generales sobre la optimización de modularidad y su fiabilidad. Un valor alto para el máximo de la modularidad, no significa necesariamente que una red cuente con una estructura modular. Las redes aleatorias también pueden tener particiones con valores de modularidad altos, a pesar de que no haya comunidades explícitamente constituidas [145, 159]. Por lo tanto, la modularidad máxima de una red revela su estructura modular sólo si es apreciablemente mayor que la modularidad máxima de una red aleatoria del mismo tamaño [242].

Además, se asume que el máximo de modularidad ofrece la "mejor" partición de la red en las comunidades. Sin embargo, esto no siempre es cierto [226, 243]. En la definición de modularidad (ecuación (3.2)) la red se compara con una versión aleatoria de ella, que mantiene los grados de sus nodos. Si los grupos de nodos en las redes están más estrechamente conectados de lo que estarían en la red aleatoria, la optimización de la modularidad los consideraría como partes de un mismo módulo. Pero si las comunidades tienen menos de $\sqrt{m}$ aristas internas, el número esperado de aristas se hay entre ellas en el modelo nulo de modularidad es menor que uno, y una sola arista de interconexión podría causar la fusión de los dos grupos en la partición óptima. Esto se mantiene para cada densidad de aristas dentro de las comunidades, incluso en el caso límite en el cual todos los nodos de cada comunidad están conectados el uno al otro, *i.e.* si los grupos son *cliques*.

Algoritmos Dinámicos.

Esta sección describe los métodos que emplean procesos dinámicos que se ejecutan en la red como interacciones de spin, sincronización y caminatas aleatorias.

Modelos de Spin.

El *modelo de Potts* es uno de los más populares de la *mecánica estadística*. En él se describe un sistema de espines que pueden estar en q estados diferentes. La interacción es ferromagnética, *i.e.* favorece la alineación espines, por lo que a temperatura cero, todos los espines están en el mismo estado. Si las interacciones antiferromagnéticas también están presentes, el estado base del sistema no puede ser aquel en el que todos los espines están alineados, sino un estado en donde los valores de los diferentes espines coexisten, en grupos homogéneos. Si variables de *spin de Potts* se asignan a los nodos de una red con estructura modular, y las interacciones son entre espines vecinos, es posible entonces que los grupos topológicos puedan ser recuperados mediante grupos de espines del sistema, ya que hay muchas más interacciones dentro de las comunidades que en el exterior.

Método de Reichardt y Bornholdt Basandose en esta idea, Reichardt y Bornholdt [160] propusieron un método para detectar comunidades que mapea una red en un modelo *q-Potts*, con interacciones de vecinos más cercanos. Cada nodo i está etiquetado por una variable de *spin* de Potts σ_i, que indica el grupo que incluye al nodo. El hamiltoniano del modelo, *i.e.* su energía, es la suma de dos términos en competencia, uno favoreciendo la alineación de espín, y otro la antialineación:

$$\mathscr{H}(\{\sigma\}) = -\sum_{i<j} J_{ij}\delta(\sigma_i,\sigma_j) = -\sum_{i<j} J(A_{ij} - \gamma p_{ij})\delta(\sigma_i,\sigma_j) \tag{3.10}$$

donde J es una constante que expresa la fuerza de acoplamiento, A_{ij} son los elementos de la matriz de adyacencia de la red, $\gamma > 0$ un parámetro que expresa la contribución relativa a la energía de las aristas existentes y faltantes, y p_{ij} es el número esperado de enlaces conectando i y j para una red de *modelo nulo* con el mismo número total de aristas m de la red considerada.

El peso relativo de estos dos términos se expresa por un parámetro γ , que normalmente se establece en el valor de la densidad de aristas de la red. El objetivo es encontrar el estado base del sistema, *i.e.*, minimizar la energía. Esto puede hacerse con el método de *templado simulado* (sección 3.3.4.3), a partir de una configuración donde los espines se asignan aleatoriamente a los nodos y el número de estados q es muy alto. El procedimiento es bastante rápido y los resultados no dependen de q. El método también permite identificar nodos compartidos entre comunidades, a partir de la comparación de las particiones correspondientes a la energía global y local mínimas. Posteriormente, Reichardt y Bornholdt obtuvieron un marco general [161], en el cual

detectar la estructura modular es equivalente a encontrar el estado base de un modelo *q-Potts* de espín de cristal (*Spin glass theory*). Su método anterior y la *optimización de modularidad* (sección 3.3.4) se recuperan como casos especiales. Se pueden descubrir comunidades superpuestas mediante la comparación de particiones con el mismo (mínimo) de energía, y la estructura jerárquica puede investigarse ajustando un parámetro que actúa sobre la densidad de las aristas de una red de referencia, sin estructura de comunidad.

Método de Ronhovde y Nussinov Otra técnica similar basada en el *modelo de Potts*, fue propuesta por Ronhovde y Nussinov [244]. La energía de su modelo de *spin* es:

$$\mathscr{H}(\{\sigma\}) = -\frac{1}{2}\sum_{i\neq j}\left[A_{ij} - \gamma A_{ij}\right]\delta(\sigma_i, \sigma_j) \tag{3.11}$$

La gran diferencia con la ecuación (3.10) es la ausencia de un término *modelo nulo*. El modelo considera pares de nodos en la misma comunidad: las aristas entre los nodos se recompensan energéticamente, mientras que las aristas faltantes se penalizan. El parámetro γ corrige el intercambio entre las dos contribuciones. La energía se minimiza mediante el desplazamiento secuencial de nodos/espines individuales a las comunidades que producen la mayor disminución de la energía del sistema, hasta la convergencia.

Label Propagation Method

Otra propuesta dinámica bastante rápida es la de Raghavan et al. [245], quienes diseñaron un método simple y rápido basado en la *propagación de etiquetas*. Los nodos reciben inicialmente etiquetas únicas. En cada iteración, se realiza un barrido sobre todos los nodos, en orden secuencial aleatorio: cada nodo toma la etiqueta compartida por la mayoría de sus vecinos. Si no hay una mayoría única, una de las etiquetas de la mayoría se elige al azar. De esta forma, las etiquetas se propagan a través de la red: la mayoría de las etiquetas desaparecerán y otras dominarán. El proceso alcanza la convergencia cuando cada nodo tiene la etiqueta de mayoría de sus vecinos. Las comunidades se definen como grupos de nodos con etiquetas idénticas en convergencia.

El algoritmo no ofrece una solución única. Debido a los muchos vínculos encontrados a lo largo del proceso, es posible derivar diferentes particiones a partir de la misma condición inicial, con diferentes semillas aleatorias. Las pruebas en redes reales muestran que todas las particiones encontradas son similares entre sí. Sin embargo, la principal ventaja del método es el hecho de que no necesita ninguna información sobre el número y el tamaño de las comunidades. No necesita ningún parámetro, tampoco. La complejidad de tiempo de cada iteración del algoritmo es $O(m)$ y el número de iteraciones a convergencia parece ser independiente del tamaño de la red, o crece muy lentamente con él. Entonces la técnica es realmente rápida y puede usarse para el análisis de sistemas grandes. Se ha demostrado que que el método es equivalente a encontrar los mínimos

energéticos locales de un *modelo de Potts* simple a temperatura cero, y que el número de mínimos de energía es considerablemente mayor que el número de nodos de la red. Así también se han propuesto mejoras al método por diferentes autores [146]. Las pruebas del algoritmo modificado en el *benchmark LFR* (sección 3.4.2.3) dan buenos resultados y fomentan futuras investigaciones.

Sincronización

La sincronización es otro proceso dinámico prometedor para revelar comunidades en redes complejas. Si se colocan osciladores en los nodos, con fases iniciales aleatorias, y tienen interacciones de vecino cercano, los osciladores de la misma comunidad se sincronizarán primero, mientras que una sincronización completa requerirá mayor tiempo. Por lo tanto, si se sigue la evolución en el tiempo del proceso, los estados con grupos sincronizados de nodos pueden ser bastante estables y de larga duración, así que pueden ser fácilmente reconocidos. Esto fue demostrado por primera vez por Arenas, Díaz-Guilera y Pérez Vicente [246], utilizando osciladores de Kuramoto [247], que son vectores bidimencionales dotados con una frecuencia de oscilación adecuada.

$$\frac{d\theta_i}{dt} = \omega_i + \sum_j K sen(\theta_j - \theta_i) \tag{3.12}$$

Si la interacción de acoplamiento excede un umbral, la dinámica conduce a la sincronización. Arenas *et al.* mostraron que la evolución temporal del sistema revela algunas escalas de tiempo intermedias correspondientes a escalas topológicas de la red, *i.e.* para diferentes niveles de organización de los nodos. De esta manera se puede revelar la estructura de comunidad jerárquica. El algoritmo escala en un tiempo $O(mn)$ y $O(n^2)$ en redes *sparse*, y da buenos resultados en ejemplos prácticos. Sin embargo, los algoritmos basados en sincronización pueden no ser confiables cuando las comunidades son de tamaño muy diferente.

Algoritmos basados en Caminatas Aleatorias

Usar caminatas aleatorias para encontrar comunidades proviene de la idea de que un caminante aleatorio en una red, pasaría mucho tiempo dentro de una comunidad debido a la alta densidad de aristas y el consecuente número de caminos que pueden seguirse.

Un primer método fue propuesto Zhou [248] quien utilizó caminatas aleatorias para definir una distancia entre pares de nodos (dado que es probable que nodos cercanos bajo está distancia pertenezcan a la misma comunidad) y definiciones globales y locales de nodos atractores para definir comunidades, las cuales son sub-redes mínimas. Las aplicaciones a redes reales y artificiales muestran que el método puede encontrar particiones significativas. Una medida diferente distancia entre nodos basada en caminatas aleatorias fue presentada por Latapy y Pons [249] en su método *Walktrap*. La distancia se calcula a partir de las probabilidades de que el caminante aleatorio se mueva de un

nodo a otro en un número de pasos fijo. Los nodos se agrupan en comunidades a través de *agrupación jerárquica* (sección 3.3.1). El método es bastante rápido, ejecutándose hasta el final en un tiempo $O(n^2 log(n))$ en una red *sparse*.

Algoritmo de Agrupamiento de Markov (Markov Cluster Algorithm - MCL). Este método, desarrollado por Van Dongen [209], simula un proceso de difusión en una red. Se parte de la matriz estocástica de la red, que se obtiene a partir de la matriz de adyacencia dividiendo cada elemento A_{ij} por el grado de i. El elemento S_{ij} de la matriz estocástica da la probabilidad de que un caminante aleatorio, estando en el nodo i, se mueva a j. La suma de los elementos de cada columna de $\mathbb{S}$ es uno.

Cada iteración del algoritmo consta de dos pasos. En el primer paso, llamado *expansión*, la matriz estocástica de la red se eleva a una potencia entera p (usualmente $p = 2$). La entrada M_{ij} de la matriz resultante da la probabilidad de que un caminante aleatorio, partiendo del nodo i, alcance a j en p pasos (difusión). El segundo paso, que no tiene contraparte física, consiste en elevar cada entrada de la matriz M a una potencia $\alpha \in \mathbb{R}$. Esta operación, llamada *inflación*, aumenta los pesos entre pares de nodos que tienen valores grandes de difusión, los cuales probablemente están en la misma comunidad. A continuación, los elementos de cada fila deben ser divididos por su suma, de tal manera que la suma de los elementos de la fila es igual a uno con lo que se recupera una nueva matriz estocástica. Después de algunas iteraciones, el proceso proporciona una matriz estable, con algunas propiedades notables. Sus elementos son cero o uno, por lo que es una especie de matriz de adyacencia. Más importante aún, la red descrita por la matriz se desconecta, y sus componentes conectados son las comunidades de la red original.

El método es muy simple de implementar, lo que es la principal razón de su éxito: hasta ahora, el *MCL* es uno de los algoritmos de agrupamiento más usados en bioinformática. Dada a la multiplicación de matrices del paso de expansión, el algoritmo debe escalar como $O(n^3)$, incluso si la red es escasa, ya que la matriz en ejecución se vuelve densa rápidamente, después de unos pocos pasos del algoritmo. Un problema de este método es el hecho de que la partición final es sensible al parámetro α utilizado en el paso de inflación. Por lo tanto se pueden obtener varias particiones, y no es claro cuales son los más significativas o representativas.

Una propuesta muy similar al *MCL* fue propuesta por Alcalá-Corona *et al.* [250] en 2011. Es un método *greedy* que usa la misma idea difusión en una red, pero sin usar el formalismo matricial. También tiene una etapa de *expansión* donde se calcula la probabilidad de alcanzar cualquier nodo en T pasos (donde usualmente $T \leq 3$). Después se implementa una agrupación en términos de un parámetro p^* de corte, proporcionado *a priori* al algoritmo. La gran desventaja es que el algoritmo es lento para redes grandes en términos del parámetro T, ejecutándose con una complejidad $O(m^T n)$ además de que depende de los parámetros T y p^*. Sin embargo ha sido probado en la *prueba LFR*

(sección 3.4.2.3) con buenos resultados[1].

Infomap y la *mapequation*. Rosvall y Bergstrom, atacaron el problema de detectar comunidades en redes combinado caminatas aleatorias y teoría de la información. Se preguntaron: ¿cuál es la mejor forma de describir una caminata aleatoria infinitamente larga en la red? El objetivo es comprimir de forma óptima la información necesaria para describir dicho proceso de difusión usando la misma caminata aleatoria como *proxy*. De ésta idea se deriva la denominada *mapequation* o *ecuación de mapa* (ecuación 3.13) [156].

La descripción más simple se obtiene al enumerar secuencialmente todos los nodos alcanzados por el caminante aleatorio, describiendo cada nodo por una *palabra código* única. Así el contenido de información está dado por el número total de bits requeridos para indicar las diversas etapas de la caminata. De está manera es posible lograr una descripción de dos niveles, en la que se dan nombres únicos a las estructuras importantes de la red y a los nodos dentro de la misma estructura. Si la red tiene estructura modular puede haber una descripción más compacta, dado que las *palabras código* asociadas a un nodo se pueden reciclar entre diferentes estructuras, que desempeñan el papel de regiones en el mapa, y los nodos con nombre idéntico se distinguen al especificar la región (comunidad) a la que pertenecen. Esto es similar al procedimiento generalmente adoptado en los mapas geográficos, donde las regiones son ciudades y generalmente hay varias ciudades y calles con el mismo nombre en todas las regiones.

La codificación de Huffman [251] se usa para nombrar nodos y asignarles su *palabra código*. Para la caminata aleatoria, las regiones antes mencionadas son las comunidades, ya que es intuitivo que el caminante pasará mucho tiempo dentro de ellas, por lo que juegan un papel crucial en el proceso de difusión de la información. La detección de comunidades en la red se convierte entonces en el siguiente problema de codificación: encontrar la partición que produce la longitud mínima de la descripción de una caminata aleatoria infinita.

Así entocnes, la *mapequation* (ecuación 3.13) proporciona la longitud de la descripción de una caminata aleatoria infinita y consta de dos términos, que expresan la entropía de Shannon [61, 62, 63, 64, 65] de la caminata dentro y entre las comunidades (regiones). Cada vez que el caminante da un paso hacia una región (comunidad) diferente, se necesita usar la palabra código de ese grupo en la descripción, para informar al decodificador que hubo una transición[2]. Claramente, si las comunidades están bien separadas unas de otras, las transiciones del caminante aleatorio entre comunidades serán poco frecuentes, por lo que es ventajoso usar el mapa, donde las comunidades son

[1] Lamentablemente el método no ha sido publicado aún.

[2] En cambio, para una descripción de un nivel, en la que todos los nodos tienen diferentes nombres, basta con especificar la palabra código del nodo alcanzada en cada paso para definir completamente el proceso.

las regiones, ya que en la descripción de la caminata aleatoria las *palabras código* de las comunidades no se repetirá muchas veces, mientras que hay un ahorro considerable en la descripción debido a la longitud limitada de las *palabras código* usadas para denotar los nodos (figura **??**). En cambio, si no hay comunidades bien definidas y/o si la partición no es representativa de la estructura modular real de la red, las transiciones entre las comunidades de la partición serán muy frecuentes y habrá poca o ninguna ganancia al usar la descripción de dos niveles del mapa. Al final, la mejor partición es la que proporciona la descripción de longitud mínima.

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} q_{\curvearrowright} H(\mathcal{P}_i) \tag{3.13}$$

La minimización de la longitud de la descripción del proceso, se lleva a cabo combinando una búsqueda *greedy* con templado simulado (sección 3.3.4.3)[1]. Así, este método, llamado *Infomap*, se puede aplicar a redes pesadas, tanto dirigidas como no dirigidas. En este último caso, el proceso de caminata aleatoria se modifica al introducir una probabilidad de teletransportación τ, para garantizar la *ergodicidad* (*i.e.* que se alcance un estado estacionario no trivial de la caminata infinita), al igual que en el algoritmo PageRank de Google [120] (ver capitulo anterior, sección 2.2.5.1).

Infomap se ha ampliado sucesivamente a la detección de la estructura modular jerárquica [253] y de comunidades superpuestas [254]. Así, *Infomap* y sus variantes generalmente devuelven particiones diferentes a los métodos basados en estructura (por ejemplo, optimización de modularidad). Esto se debe a que se basa en flujos que se ejecutan en el sistema, en oposición a variables estructurales topológicas como número de aristas, nodos hub, etc. La diferencia es particularmente notable en redes dirigidas [156], donde las direcciones de las aristas restringen fuertemente los posibles flujos. Las características estructurales, obviamente, juegan un papel importante en la dinámica de los procesos que se ejecutan en redes, pero la dinámica no puede reducirse generalmente a una interacción de elementos estructurales, al menos no simples como, por ejemplo, el grado de los nodos. Sin embargo, a veces, los enfoques estructurales y dinámicos son equivalentes. La minimización de la longitud de la descripción de la caminata $L(M)$ de *Infomap* es equivalente e incluso mejor que maximizar la modularidad Q de Newman-Girvan (sección 3.3.4). Se profundizara más a detalle sobre *Infomap* en la sección 4.1.1 del siguiente capítulo.

[1] En un artículo sucesivo [252], los autores adoptaron la técnica rápida *greddy* de Blondel et al. para la optimización de la modularidad [162] (sección 3.3.4.2), con algunos refinamientos.

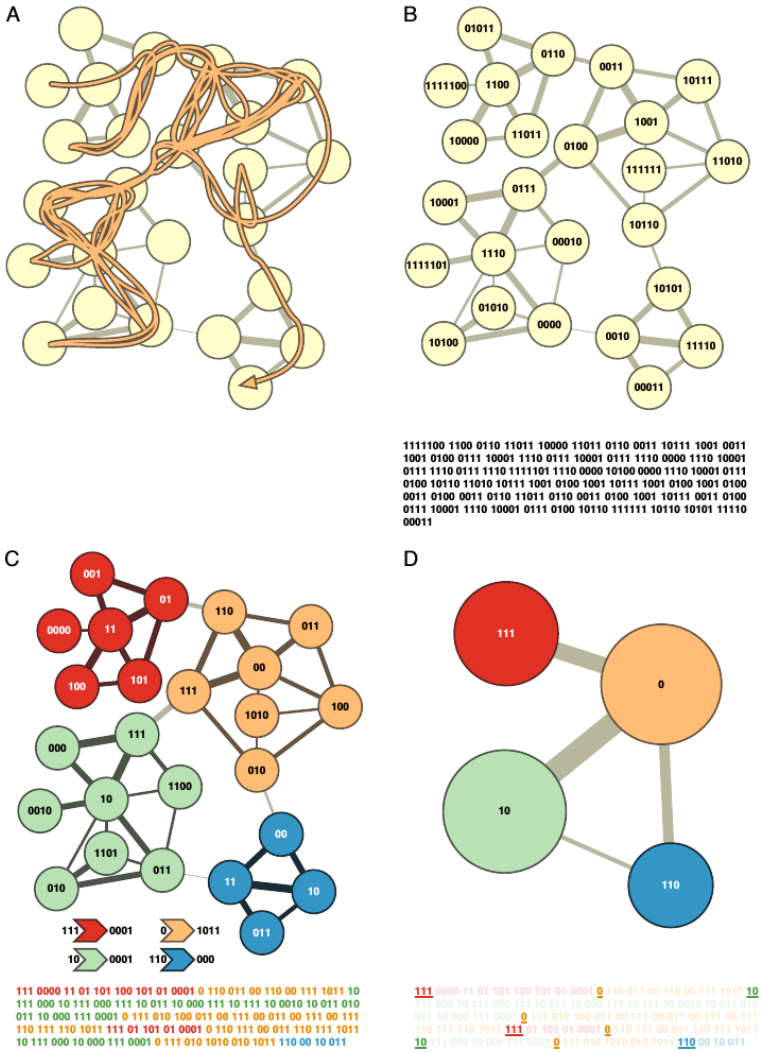


Figura 3.13: **Infomap de Rosvall y Bergstrom**. La caminata aleatoria en (**A**) se puede describir como una secuencia de nodos, cada uno etiquetado con *palabras de código* únicas (**B**), o dividiendo la red en regiones y usando *palabras de código únicas* solo para los nodos de la misma región (**C**). De esta forma, la misma *palabra de código* puede usarse para múltiples nodos, a costa de indicar cuándo el caminante aleatorio abandona una región para ingresar a una nueva, ya que en ese caso hay que especificar la palabra de código de la nueva región para ubicar al caminante. La red tiene cuatro comunidades (indicadas por los colores en (C)), y en este caso la descripción tipo mapa de (C) es más compacta que la de (B). Esto se muestra al mirar el código real que se necesita en cualquier caso (parte inferior de las figuras), que es claramente más corto para (C). En (**D**) las transiciones entre los módulos están resaltadas en la *codificación*.

Otras metodologías

OSLOM

Lancichinetti *et. al.* con su *Método de Optimización Local de Estadísticas de Orden* (*Order Statistics Local Optimization Method* - OSLOM) [255], encuentran comunidades a través de la optimización local de un *score* de significancia estadistica. Este *score* se basa en la idea de que que no es suficiente identificar grupos en la red, sino también hay que preguntarse qué tan significativos o no aleatorios son. Desafortunadamente, la mayoría de los algoritmos no pueden evaluar la importancia de sus resultados.

Si las comunidades detectadas son compatibles con fluctuaciones aleatorias, no son grupos adecuados y deben descartarse. Mientras menor sea la probabilidad de que se generen por aleatoriedad, más seguros podemos estar de que las comunidades reflejan una estructura modular real. Lo anterior se puede lograr si se cuenta con un *modelo nulo* confiable[1] que describa cómo se puede *aleatorizar* la estructura de la red en estudio y que permita estimar qué tan probable es que la estructura del grupo candidato se genere de esta manera. Un *score* de significancia se puede estimar por ejemplo a parir de la significancia estadística de la medida de modularidad Q_{max}. El cual está dado por la distancia de Q_{max} desde el promedio nulo del modelo $\langle Q_{rand} \rangle$ en unidades de la desviación estándar σ_Q^{rand}, es decir, por el *z-score*

$$z = \frac{Q_{max} - \langle Q_{rand} \rangle}{\sigma_Q^{rand}} \tag{3.14}$$

Si $z \gg 1$, entonces Q_{max} indica una fuerte estructura modular. Así OSLOM se basa en este principio para estimar un *score* de significancia para la estructura modular al calcular varias propiedades de las comunidades, como su densidad interna, y compararlos con los valores del *modelo nulo*.

Stochastic Block Models (SBM)

La inferencia estadística proporciona un poderoso conjunto de herramientas para abordar el problema de la detección de comunidades. El enfoque estándar es ajustar un *modelo generador de red (generative network model)* a los datos. El **modelo estocástico de bloques** (SBM, por sus siglas en inglés) es, por mucho, el modelo generador más utilizado de redes con estructura modular. Además de que puede describir otros tipos de estructura de grupo en redes (figura 3.14). La máxima probabilidad logarítmica no normalizada de que una partición g dada, en q comunidades de la red G sea reproducida por el SBM estándar [258] es:

$$\mathscr{L}_S\left(G|g\right) = \sum_{r,s=1}^{q} e_{rs}\, log\left(\frac{e_{rs}}{n_r n_s}\right) \tag{3.15}$$

[1] El modelo de configuración [256, 257] es un *modelo nulo* popular en la literatura

donde e_{rs} es el número de aristas que van de la comunidad r a la comunidad s, n_r (n_s) es el número de nodos en r (o s) y la suma se extiende sobre todos los pares de comunidades (incluso cuando $r = s$).

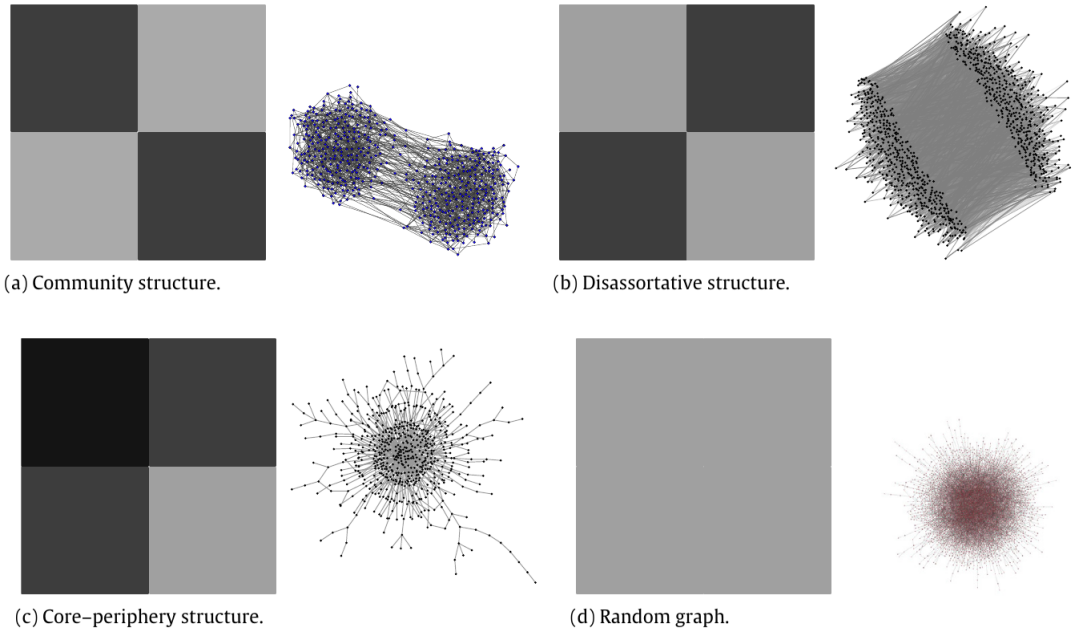


Figura 3.14: **Stochastic Block Model**. Se muestran las matrices de adyacencia esquemáticas de redes generadas por el modelo para elecciones especiales de probabilidades de arista. Los bloques más oscuros indican mayores probabilidades de arista y, en consecuencia, una mayor densidad de aristas dentro del bloque. (a) Ilustra la estructura modular (o assortative): las probabilidades (densidades de enlace) son mucho más altas dentro de los bloques diagonales que en cualquier otro lugar. (b) Muestra la situación opuesta (estructura disassortativa). (c) Ilustra una estructura núcleo-periferia. (d) Muestra un gráfico aleatorio tipo Erdős y Rényi: todas las probabilidades de arista son idénticas, dentro y entre los bloques, por lo que no hay grupos reales.

El inconveniente más importante de este tipo de enfoques es la necesidad de especificar el número q de comunidades de antemano, que generalmente es desconocido para las redes reales. Sin embargo en un reciente método de Tiago P. Peixoto [259], presenta una formulación bayesiana de modelos de bloques estocásticos (SBM) pesados que pueden utilizarse para inferir la estructura modular a gran escala de las redes pesadas, incluida su organización jerárquica. Dado que el método no es paramétrico, no requiere el cono-

cimiento previo de la cantidad de comunidades u otras dimensiones del modelo, que en cambio se deducen de los datos. La aplicación del método incluye varias redes pesadas empíricas, como redes de migración, patrones de votación parlamentario y conexiones neuronales en el cerebro humano. Finalmente se ha mostrado que este tipo de enfoques y el de maximización de modularidad Q (sección 3.3.4) son equivalentes [260].

Percolación de Clique (Clique Percolation).

Un método que toma en cuenta tanto la localidad de la definición comunidad como la posibilidad de tener comunidades superpuestas es el método de Percolación de Clique (*Clique Percolation Metod*) CPM por sus siglas en inglés, propuesto por Palla *et. al.* [175]. Se basa en el concepto de que las aristas interiores de la comunidad son propensas a formar *cliques*[1] debido a su alta densidad. Por otro lado, es poco probable que las *aristas inter-comunidades* formen *cliques* [2].

Palla et al. definen un *k-clique* como un red completa con k nodos. Si fuera posible que un *clique* se "moviera" en una red de alguna manera, probablemente se quedaría atrapado dentro de su comunidad original, ya que no podía cruzar el cuello de botella formado por las aristas entre las comunidades. Así, una comunidad *k-clique* es la sub-red conectada más grande obtenida por la unión de un *k-clique* y de todos los *k-cliques* que están conectados a la misma. Se podría decir que una comunidad *k-clique* se identifica haciendo "rodar" un *k-clique* sobre los *k-cliques* adyacentes, donde "rodar" significa rotar de un *k-clique* sobre los $k\breve{}1$ nodos que comparte con cualquier *k-clique* adyacente. Por construcción, las comunidades *k-clique* pueden compartir nodos, para que puedan superponerse.

Para encontrar comunidades *k-clique*, primero se buscan *cliques* máximos, una tarea que requiere un tiempo de ejecución que crece exponencialmente con el tamaño de la red. Sin embargo, los autores encontraron que para las redes reales que ellos analizaron, el procedimiento es bastante rápido, permitiendo analizar redes con hasta 10^5 nodos en un tiempo razonablemente corto. El *CPM* tiene el mismo límite que el algoritmo de Radicchi *et. al.*: se asume que la red tiene un gran número de cliques, por lo que puede fallar para proporcionar particiones significativas en redes con sólo unos pocos cliques.

Métodos de prueba de algoritmos de detección de comunidades.

Como podemos ver, existen una gran variedad de métodos, técnicas y algoritmos para detectar comunidades en redes complejas. Sin embargo, una cuestión importante

[1] Pequeñas sub-redes completas.

[2] esta idea también fue utilizada en el método divisivo de Radicchi *et. al.* (sección 3.3.2.2)

en este campo es que a pesar de existir muchísimos métodos es difícil comparar su rendimiento y exactitud. A menudo se confía en algunos algoritmos, por razones que tal vez no tienen que ver con el desempeño real de los mismos, como por ejemplo la popularidad del método o de su inventor. Muchos algoritmos no son probados o comparados con otros para medir su eficiencia, ya que es difícil establecer si un método es confiable o tiene mejores resultados respecto de otro. Comparar el rendimiento de algoritmos es una tarea que había recibido poca atención en la literatura y que se ha abordado en los últimos años [146]. Probablemente una razón por la cual los algoritmos y técnicas de a de detección de comunidades proliferaron tanto en los últimos años, es que no se contaba con un conjunto de pruebas estándar, en el que hubiese consenso por parte de la comunidad que estudia el tema.

Así, para probar un nuevo algoritmo de detección de comunidades sería deseable poder tener una red en la cual las comunidades (módulos) estén perfectamente establecidas y se conozcan *a priori*, para comprobar si el algoritmo es capaz de encontrarlas, o que tanto se acerca a hacerlo. De esta manera se tendría un problema con una solución bien definida, y se podrían comparar métodos. Sin embargo, esto no es trivial, pues muchos algoritmos se basan en nociones intuitivas (muy similares) de lo que es una comunidad, pero con diferentes implementaciones.

A pesar de ésta problemática, existen en general dos formas en las que se prueban los algoritmos publicados en la literatura: en *redes reales*, que son un grupo reducido que existen desde hace tiempo, y *redes generadas computacionalmente*, para las cuales existen varias propuestas dentro de la literatura y podrían constituir un conjunto de redes de referencia (benchmarks) más fiables para probar los métodos y algoritmos. Sin embargo, cabe señalar que es crucial que quienes estudian el tema estén de acuerdo en un método de prueba estandarizado. A continuación se presentan algunas de las pruebas más utilizadas en la literatura tanto para el caso de redes reales como para redes generadas por computadora.

Redes Reales

Para el caso de las redes reales, las comunidades son conocidas mediante estudios no topológicos del propio autor de la red. Dado el estudio de los fenómenos que dan origen a estas redes, se cuenta con información precisa acerca de los nodos y sus propiedades. Así, podemos encontrar varias redes que se usan regularmente como pruebas para algoritmos de detección de comunidades tales son los casos de la *red del Club de Karate de Zachray* (*Zachary's karate ckub*) [149], la *red de delfines de nariz de botella de Lusseau* [150], la *red de equipos de fútbol americano colegial* de universidades estadounidenses [157], entre otras. Sin embargo a menudo los módulos o comunidades en estas redes se construyen de manera subjetiva con base en la información del estudio hecho por el autor, lo que en ocasiones puede no ser la mejor opción para probar un algoritmo.

Cabe señalar que cuando se trata de redes reales, es útil resolver su estructura modular con diferentes técnicas, para cotejar los resultados y asegurarse de que sean compatibles entre sí, ya que en algunos casos la respuesta puede depender fuertemente del algoritmo específico adoptado. Sin embargo, hay que tener en cuenta que no hay ninguna garantía de que las comunidades "razonables", definidas a partir de *información no estructural* deban coincidir con las detectadas por los métodos basados exclusivamente en la estructura de la red.

Red del Club de Karate de Zachary.

La red real más popular con una estructura modular conocida y por mucho la más estudiada, es la red social del **Club de Karate de Zachary** (ver sección 3.1.1.1). Esta es una red social que representa las relaciones personales entre los miembros de un club de karate en una universidad estadounidense (figura 3.15). Durante dos años, el sociólogo Wayne Zachary observó los vínculos entre los miembros, tanto adentro como afuera del club [149]. En algún momento, surgió un conflicto entre el administrador del club (nodo 1) y uno de los maestros (nodo 33), lo que llevó a la división del club en dos clubes más pequeños, algunos miembros se quedaron con el administrador y los otros siguieron al instructor. Los nodos de las dos comunidades se diferencian con cuadrados y círculos en la figura 3.15.

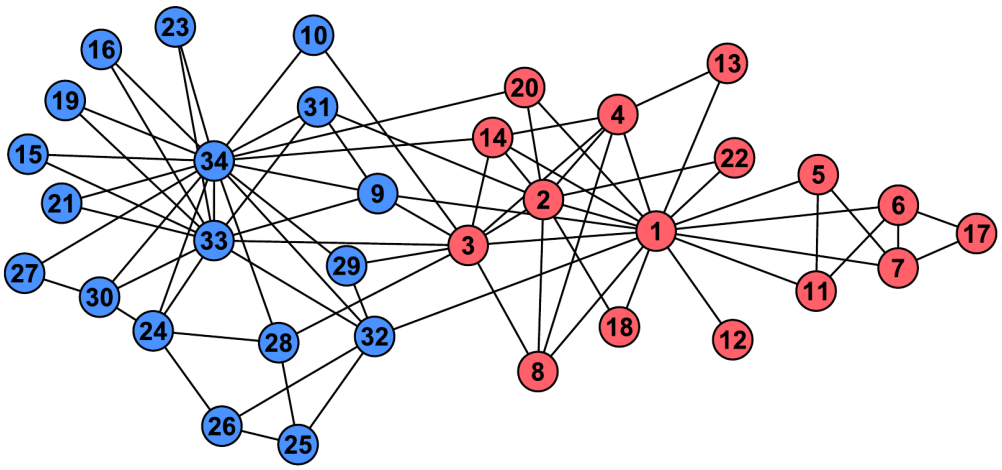


Figura 3.15: **Red del *Club de Karate* de Zachary**. Se muestra la partición estándar en dos comunidades. Los círculos rojos representan una comunidad y los azules la otra.

Aquí, la pregunta es si la separación real en dos grupos sociales pueden predecirse a partir de la topología de la red. Varios algoritmos, de hecho, son realmente capaces de identificar las dos comunidades sin contar algunos nodos intermedios, que pueden ser

clasificados erróneamente, por ejemplo los nodos 3 y 10). Otros métodos son menos exitosos y regresan a una división de la red en cuatro grupos. Sin embargo, es fundamental subrayar que la comparación de las comunidades detectadas por los distintos métodos con la división del *club de karate de Zachary* se basa en un supuesto muy fuerte: que la división en realidad reproduce la separación de la red social en dos comunidades. No hay ningún argumento real, más allá del sentido común, que apoye a esta suposición.

Otras Redes Reales.

Red de equipos de fútbol americano colegial. Otro ejemplo que se han utilizado con frecuencia en la literatura para probar algoritmos de detección de comunidades es la red de equipos de fútbol americano colegial de universidades estadounidenses (figura 3.17) obtenida por Girvan y Newman [157], en ésta hay 115 nodos, que representan a los equipos, y dos nodos están conectados si sus equipos juegan unos contra otros.

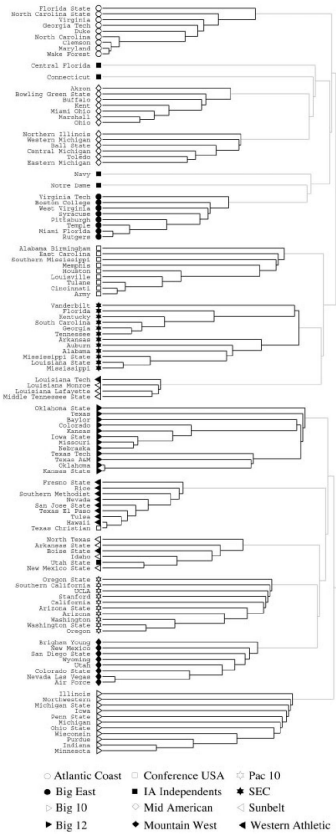


Figura 3.16: **Red de equipos de fútbol americano colegial de universidades estadounidenses.**

Los equipos se dividen en 12 conferencias. Los partidos entre equipos de la misma conferencia son más frecuentes que los partidos entre equipos de diferentes conferencias, así que se tiene una partición natural en el que las comunidades corresponden a las conferencias.

Red de delfines de nariz de botella de Lusseau. Finalmente, tenemos la red social de los delfines nariz de botella (ver sección 3.1.1.1) que viven en Doubtful Sound (Nueva Zelanda), estudiada y construida por Lusseau [150] (figura 3.17)

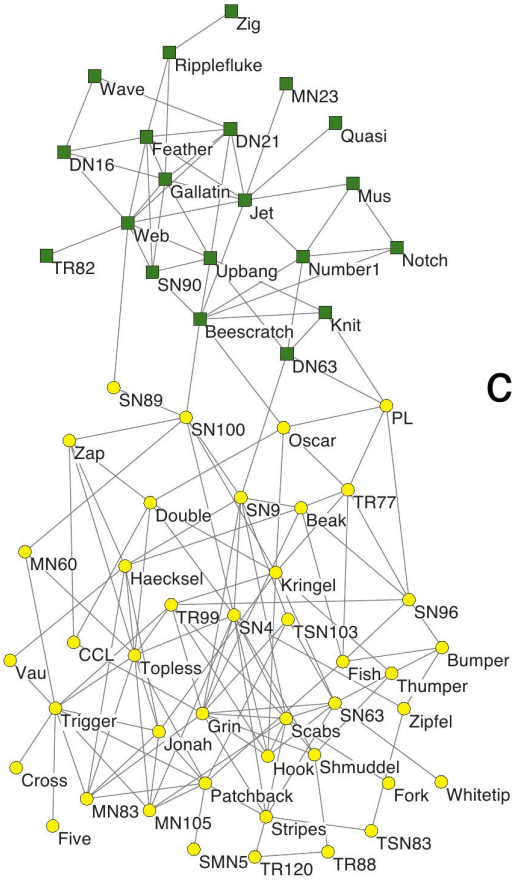


Figura 3.17: **Red de delfines de naris de botella estuadiada por Lusseau**. En cuadros de color verde se representa una comunidad y los circulos amarillos la otra.

También para estas redes aplica la misma salvedad: nada garantiza que las comunidades "razonables", definidas sobre la base de la información no topológica, deban coincidir con los detectadas por los métodos basados exclusivamente en la topología de la red.

Redes de prueba (benchmarks) generadas computacionalmente.

Otra propuesta para probar algoritmos de detección de comunidades es el caso de las redes generadas por computadora, donde tanto la red como sus comunidades pueden ser diseñadas *ad hoc*. Procediendo de esta manera tanto la red y como su estructura modular son conocidos a priori y están bien determinadas. Por ésta razón los *benchmarks* generados computacionalmente pueden ser más fiables para probar métodos y algoritmos que incorporan la *estructura a gran escala* de las redes es su metodología.

Existen varias propuestas para generar redes de forma computacional, pero nos centraremos en una clase especial de éstas que se generan con el denominado modelo de **partición-l plantada** (*planted l-partition model*) [261]. En particular existen dos pruebas desarrolladas y exploradas ampliamente en la literatura, la *prueba GN* [157] y la *LFR* [262], entre otros benchmarks [146].

Modelo de *partición-l plantada.*

El denominado modelo de partición-l plantada (*planted l-partition model*) [261] particiona una red con $n = g \cdot l$ nodos en l grupos con g nodos cada uno. Los nodos de un mismo grupo (comunidad) están vinculados con una probabilidad p_{in} fija, mientras que los nodos de grupos diferentes están vinculados con una probabilidad p_{out}. Cada módulo corresponde entonces un *grafo aleatorio* tipo Erdös-Rényi con probabilidad de conexión $p = p_{in}$, asimismo si cada comunidad fuera un nodo, toda la red también sería una red tipo Erdös-Rényi con $p = p_{out}$ (véase sección 2.3.1).

Entonces para una sub-red que representa una comunidad C tendríamos que $\langle k \rangle_{in} = p_{in}(g-1)$. Y el grado promedio externo es el resto de conexiones posibles $\langle k \rangle_{out} = g \cdot p_{out}(l-1)$[1]. De ésta manera el grado promedio de la red es:

$$\langle k \rangle = p_{in}(g-1) + g \cdot p_{out}(l-1) \tag{3.16}$$

Así entonces, si $\langle k \rangle_{in} > \langle k \rangle_{out}$, es decir, si el grado promedio *intra-comunidad* excede el grado promedio *inter-comunidad* entonces la red tendrá estructura modular. Lo anterior se puede lograr variando p_{in} y p_{out} de tal manera que $p_{in} > p_{out}$, sin embargo si el grado promedio $\langle k \rangle$ es un valor fijo, p_{in} y p_{out} no son variables independientes, por lo que variar la diferencia entre el grado $\langle k \rangle_{in}$ (o externo $\langle k \rangle_{out}$) se puede lograr en términos de un solo parámetro

A continuación expondremos dos tipos de redes generadas por computadora basadas en el *modelo de partición-l plantada*. Una es la popular es prueba Girvan-Newman (*GN*

[1] Recordemos que en una red aleatoria de Erdös-Rényi con probabilidad p, el grado promedio es $\langle k \rangle = p(n-1)$ (véase sección 2.3.1).

benchmark) y la *LFR benchmark* (prueba de Lancichinetti–Fortunato–Radicchi). En este último caso la topología y estructura modular generadas son más parecidas a las redes reales descritas en el capitulo 2 y a las redes biológicas de *gran escala* expuestas en el capitulo 1, ya que la implementación del *modelo de partición-l plantada* produce redes de libre escala y el tamaño de las comunidades sigue una ley de potencias.

Prueba Girvan-Newman.

La prueba (benchmark) GN fue diseñada por Girvan y Newman [157], para probar su algoritmo y es un caso particular del *planted l-partition model*. Los autores fijan $l = 4$ y $g = 32$, de tal manera que cada red prueba se compone de 128 nodos dispuestos en 4 comunidades con 32 nodos cada uno y el grado promedio de la red se establece en $\langle k \rangle = 16$ (figura 3.18).

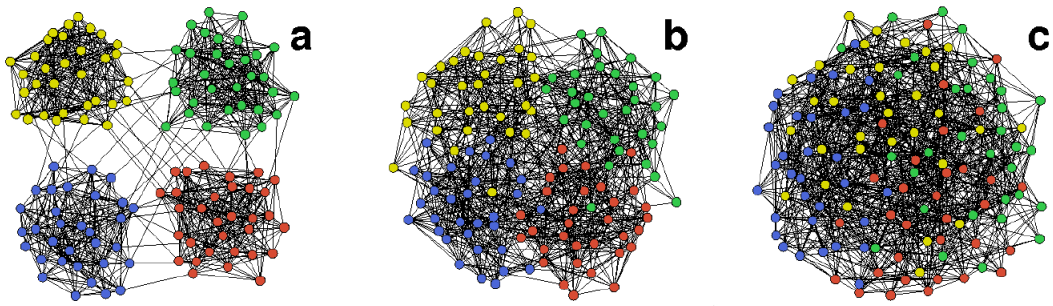


Figura 3.18: **Benchmark de Girvan y Newman**. Tres redes generadas con la *prueba GN* que corresponden a: (**a**) $\langle k \rangle_{in} = 15$; (**b**) $\langle k \rangle_{in} = 11$ y (**c**) $\langle k \rangle_{in} = 8$. Nótese que en el caso (c) los cuatro módulos son difíciles de notar.

La densidad de aristas dentro de las comunidades se ajusta variando el grado $\langle k \rangle_{in}$ y $\langle k \rangle_{out}$, de tal manera que es posible elegir valores específicos para cambiar la estructura modular de cada red prueba asegurando que $\langle k \rangle = \langle k \rangle_{in} + \langle k \rangle_{out} = 16$. Donde $\langle k \rangle_{in} = p_{in}(g-1) = 31p_{in}$ y $\langle k \rangle_{out} = g \cdot p_{out}(l-1) = 96p_{out}$. Así, se obtiene una restricción para el modelo: $p_{in} + 3p_{out} \simeq \frac{1}{2}$, pues $\langle k \rangle = 16$ y variando p_{in} y p_{out} bajo esta restricción, se pueden construir redes con una estructura modular mas fuerte o débil.

Cada nodo compartirá una fracción $\langle k \rangle_{in}$ de sus conexiones con miembros de su misma comunidad y una fracción $\langle k \rangle_{out}$ de sus conexiones con miembros de las otras comunidades. Cuando $\langle k \rangle_{in}$ es cercano a 16 (*i.e.* $p_{in} \simeq 0.5$), hay una clara estructura de comunidades, pues la mayoría de las aristas unen nodos de la misma comunidad. Y por el contrario cuando $\langle k \rangle_{in} \simeq 8$ (*i.e.* $p_{in} \simeq 0.25$) entonces $\langle k \rangle_{out} \simeq 8$ ($p_{out} \simeq 0.083$), y

entonces no hay comunidades dado que la fracción de aristas *intra-comunidad* es igual a la fracción de aristas *inter-comunidad*. En general las comunidades están bien definidas cuando $\langle k \rangle_{in} > 8$.

La facilidad de poder variar la estructura modular en términos de un solo prarámetro aunado a la simpleza de la red le ha dado a la prueba Girvan-Newman un estatus de *benchmark*. Probar un método contra el *benchmark GN* consiste en comparar mediante alguna **medida de similitud** (ver sección 3.4.3) la partición dada por el algoritmo contra la partición natural de la red en cuatro grupos del mismo tamaño. Para esto, se construyen muchas versiones diferentes de la prueba para un valor fijo de $\langle k \rangle_{in}$ y se prueba el algoritmo en cada una de ellas midiendo la similitud entre la partición proporcionada por el método y las comunidades reales de la red, promediando los resultado sobre el numero de pruebas. Lo anterior se realiza varias veces cambiando el valor de de $\langle k \rangle_{in}$, con lo que se puede determinar la eficiencia del método en redes con una estructura modular mas fuerte o débil. Los resultados se representan graficando la similitud promedio de la particion encontrada por el método contra $\langle k \rangle_{in}$. Lo cual es muy útil para poder comaprar el rendimeinto de diferentes algoritmos entre sí. La mayoría de los algoritmos por lo general hacen un buen trabajo para valores altos de $\langle k \rangle_{in}$ y comienzan a fallar cuando $\langle k \rangle_{in}$ se aproxima a 8.

Finalmente, hay que señalar que la *prueba GN* tiene un par de inconvenientes: (i) todos los nodos tienen el mismo grado esperado y (ii) todas las comunidades tienen el mismo tamaño. Estas características no son realistas, pues en general en las *redes complejas* reales se caracterizan por una distribución de grado y tamaños comunidades heterogéneos.

Prueba LFR.

Como hemos visto, el *modelo plantado de partición-l* genera redes aleatorias mutuamente interconectadas tipo Erdös-Renyi. Por lo tanto, todos los nodos tienen aproximadamente el mismo grado y las comunidades tienen exactamente el mismo tamaño por construcción. Estas dos características no concuerdan con lo que se observa en redes que representan sistemas reales, en éstas, las distribuciones de grado suelen ser asimétricas, con muchos nodos con grado bajo conviviendo con algunos nodos con grado alto. Una heterogeneidad similar se observa también en la distribución de los tamaños de las comunidades. Por lo tanto, *modelo plantado de partición-l* no es la mejor descripción de una red real con la estructura modular. Sin embargo, el modelo puede ser modificado para tener en cuenta la heterogeneidad de grados y tamaños de comunidad.

Tal es el caso de la propuesta realizada más por Lancichinetti *et al.*: el **Benchmark LFR** (Lancichinetti et al., 2008 [262]). En ésta se asume que las distribuciones de tamaño de grado y los tamaños de las comunidades suguen leyes de potencia, con exponentes τ_1 y τ_2, respectivamente. Cada nodo comparte, una fracción $1 - \mu$ de sus

aristas con los otros nodos de su comunidad y una fracción μ con los nodos de las otras comunidades, $0 \leq \mu \leq 1$ es el *parámetro de mezcla*. Las redes se construyen de la manera siguiente:

1. Se obtiene una secuencia de tamaños de comunidad (numero de nodos en cada comunidad) siguiendo la ley de distribución de potencia prescrita. Esto se hace mediante la selección de números aleatorios a partir de una ley de distribución de potencia con exponente τ_2.
2. Cada nodo i de una comunidad recibe un grado interno $(1-\mu)k_i$, donde k_i es el grado del nodo i, que ha sido tomado de una distribución de ley de potencias con exponente τ_1.
3. Todos las conexiones de los nodos de una misma comunidad son unidos entre sí aleatoriamente, hasta completar las $(1-\mu)k_i$ conexiones. De esta manera se construye la secuencia de grados internos de cada nodo en su comunidad.
4. Cada nodo i recibe ahora un número adicional de posibles conexiones, igual a μk_i (de modo que el grado final del nodo es k_i), que se conecta al azar a los nodos de otras diferentes comunidades hasta completar las k_i conexiones de cada nodo i.

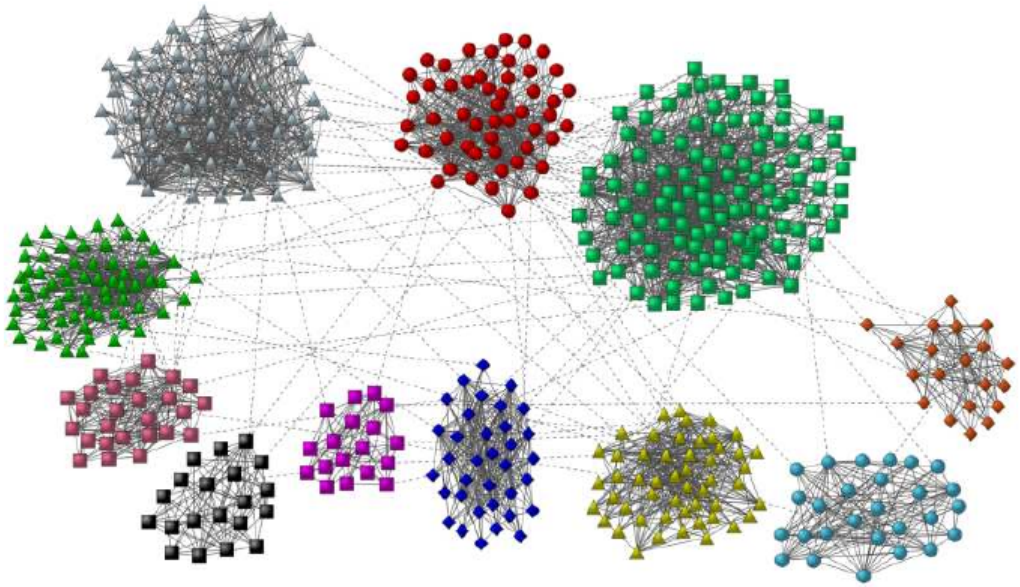


Figura 3.19: **Ejemplo de una red con 500 nodos generada con la *prueba LFR***. Las distribuciones de grado y del tamaño de comunidad siguen leyes de potencia. Este tipo de pruebas se aproximan más fielmente a las redes de mundo real con estructura modular.

De esta manera se pueden crear redes de tamaños que abarcan varios órdenes de magnitud relativamente rápido, las pruebas numéricas muestran una complejidad $O(m)$. La figura 3.19 muestra un ejemplo de una red *benchmark LFR*. La *prueba LFR* se ha ampliado para redes dirigidas y pesadas, así como a redes con comunidades superpuestas [263, 264].

Medidas de comparación de particiones.

Cuando un algoritmo divide una red en comunidades, éste realizó una partición de la red en sub-redes más pequeñas. Ahora bien, si se quiere comprobar el rendimiento de un algoritmo en redes muy grandes (en especial en redes generadas computacionalmente), es necesario establecer o definir un criterio claro para poder comparar que tan "similares" son las comunidades encontradas (partición encontrada) por el algoritmo respecto a la partición determinada por la red *benchmark* (que se desea recuperar) y que se conoce *a priori*. Podemos plantear el problema de las siguiente manera:

Sean $X = (X_1, X_2, \ldots, X_{n_X})$ y $Y = (Y_1, Y_2, \ldots, Y_{n_Y})$ dos particiones de una red G, con n_X y n_Y comunidades respectivamente. Y sea $\mathbf{n_{ij}}$ el número de nodos compartidos por las comunidades X_i y Y_j, es decidir, $n_{ij} = |X_i \cap Y_j|$. Entonces, la pregunta se puede replantear como: ¿de qué manera podemos cuantificar cuándo n_{ij} es máximo al comparar las comunidades X_i con las Y_j?, es decir, ¿cuándo es lo más similar posible la partición X a la partición Y? Para resolver esto en la literatura se han propuesto varias medidas para la similitud de particiones.

Fracción de nodos correctamente clasificados.

La *fracción de nodos correctamente clasificados* fue propuesta por Girvan y Newman para probar su algoritmo usando el *benchmark GN* [157]. El criterio para decidir la correcta clasificación de un nodo es si está en la misma comunidad con al menos la mitad de sus compañeros "naturales".

Se parte de un tener una partición A de la red, para compararla con la partición *real* B y se procede de la siguiente manera: se ubica a los conjuntos A_i más grandes encontrados por el algoritmo, que estén formados por nodos de las cuatro comunidades "reales" B_i conocidas a priori. Si mas de la mitad de los nodos de alguno de estos conjuntos grandes A_i son los mismos que los de B_i , entonces los nodos de A_i se consideran clasificados correctamente. En caso contrario si son menos de la mitad (o bien si estos conjuntos A_i nos son tan grandes como para rebasar la mitad de un conjunto real B_i), entonces todos los nodos en de estos conjuntos A_i se consideran incorrectamente clasificadas. Todos los otros nodos que no están en alguno de los conjuntos grandes A_i, también se consideran clasificados incorrectamente.

Esta manera de etiquetar los nodos como clasificados *correcta* o *incorrectamente* es un criterio muy duro y se puede considerar algo arbitrario. Hay casos en los que se podría considerar que algunos nodos han sido identificados correctamente, pero éste criterio no lo hace. Al final del proceso se cuentan cuantos nodos fueron clasificados correctamente y se divide por el tamaño total de la red, para producir un número entre 0 y 1. La medida se puede generalizar para *redes benchmark* diferentes a la de Girvan-Newman.

Información Mutua Normalizada.

Una *medida de similitud* mucho menos arbitraria se basa en reformular el problema de comparar particiones como un problema de decodificación de mensajes en el marco de la teoría de la información [61, 62, 63, 64, 65, 66]. La idea es que, si dos particiones son similares, se necesita muy poca información para inferir una partición dada la otra. La información adicional no necesaria se puede utilizar como una medida de la disimilitud.

$$I(X,Y) = \sum_x \sum_y P(x,y)\ log \frac{P(x,y)}{P(x)P(y)} \tag{3.17}$$

Esta cantidad mide cuánta información obtenemos de X si sabemos Y, y viceversa, y se puede aplicar también a las particiones X e Y, ya que son descritas por variables aleatorias. De esta manera $I(X,Y) = H(X) - H(X|Y)$, donde $H(X) = -P(x)logP(x)$ es la entropía de Shannon de X y $H(X|Y) = -P(x,y)\ logP(x|y)$ es la entropía condicional de X dado Y.

La información mutua no es una medida de similitud ideal: de hecho, todas (o algunas de) las particiones derivadas de una partición X mediante una partición adicional tendrían la misma información mutua con X, a pesar de que podrían ser muy diferentes unas de otras. En este caso, la información mutua simplemente sería igual a la entropía $H(X)$, pues la entropía condicional sería sistemáticamente cero. Para evitar esto, *Danon et. al.* adoptaron la información mutua normalizada [265]:

$$I_N(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)} \tag{3.18}$$

Ésta, también se puede calcular construyendo una matriz de confusión $\mathbb{N}$, donde las filas corresponden a las comunidades "reales", y las columnas corresponden a las comunidades "encontradas". El elemento N_{ij} de $\mathbb{N}$, es el número de nodos en la comunidad "real" i (conocida *a priori*) que también están en la comunidad j "encontrada" (detectada por el método a probar). Dado que las particiones a ser comparadas pueden tener diferente número de grupos, $\mathbb{N}$ no suele ser una matriz cuadrada. De esta manera, la similitud de dos particiones A y B está dada por la siguiente expresión:

$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\ log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A} N_{i.}\ log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{C_B} N_{.j}\ log\left(\frac{N_{.j}}{N}\right)} \tag{3.19}$$

donde el número de comunidades reales (en la partición A) se denota por C_A y el número de comunidades encontradas (en la partición B) se denota por C_B, la suma sobre la fila i de la matriz N_{ij} se denota como $N_{i.}$ y la suma sobre la columna j se denota como $N_{.j}$. Si las particiones encontradas son idénticas a las comunidades reales, entonces $I(A,B)$ toma su valor máximo de 1. Si la partición encontrada por el algoritmo es totalmente independiente de la partición real, por ejemplo, cuando se encuentra que toda la red es una sola comunidad, entonces $I(A,B) = 0$.

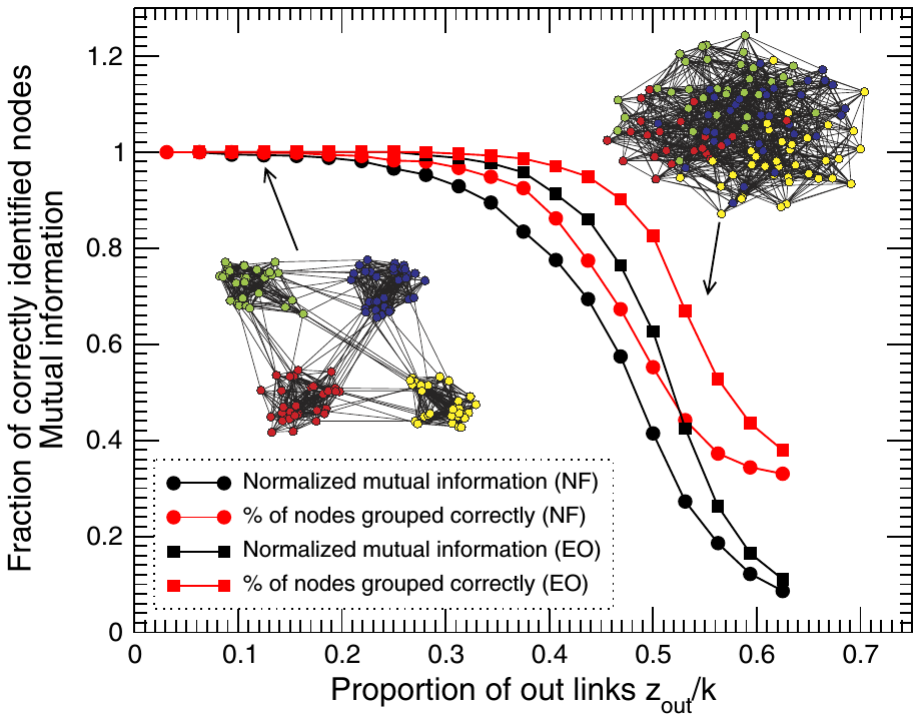


Figura 3.20: **Ejemplo de una red con 500 nodos generada con la *prueba LFR*.** Las distribuciones de grado y del tamaño de comunidad siguen leyes de potencia. Este tipo de pruebas se aproximan más fielmente a las redes de mundo real con estructura modular.

Esta medida es utilizada muy a menudo en las pruebas de algoritmos de detección de comunidades, pues tiene la ventaja de que es muy sensible al rendimiento del método a probar, ya que mide la cantidad de información extraída correctamente por el algoritmo de forma explícita. Por ejemplo, si se usa una prueba basada en el *modelo plantado de partición-l* (descrito en la sección 3.4.2.1), cuando $\langle k \rangle_{out}$ es pequeño, y dos comunidades reales son agrupadas juntas por el algoritmo, esta medida no lo penaliza tan severamente, pues tiene en cuenta la capacidad de extraer al menos algo de información acerca de la estructura modular. Por otro lado, para $\langle k \rangle_{out}$ grande es capaz de detectar si las comunidades encontradas por el algoritmo tienen poco que ver con las comunidades reales, pues $I(A, B) \to 0$.

Ésta medida, definida en principio para particiones estándar, en el que cada nodo pertenece a una sola comunidad, se ha ampliado por Lancichinetti*et al.* [264] para el caso de superposición de comunidades. La extensión no es sencilla pues ahora las clasificaciones en una comunidad de cierta partición son definidas por una variable aleatoria vectorial, ya que cada nodo puede pertenecer a más de una sola comunidad al mismo tiempo.

Otras medidas de similitud. Existen más medias de similitud basadas teoría de la información que es posible usar para poder comparar la partición que realiza un algoritmo en comunidades con las definidas por la prueba (*benchmark*) [146]. Asimismo existen otras medidas de similitud que se puede dividir en categorías como: las *medidas basadas en el recuento de par* que dependen del número de pares de nodos que están clasificados en el mismo (o diferente) grupo en las dos particiones, entre la que podemos encontrar, por ejemplo, el índice de Rand o el índice de Jaccard. Y así también las *medidas basadas en coincidencia de grupo* que tienen como objetivo detectar las coincidencias más grandes entre pares de grupos de particiones diferentes, por ejemplo, la métrica normalizada de Van Dongen [209].

Comparación de algoritmos

Ahora que hemos expuesto los métodos y como compararlos. Podemos elegir cuál de ellos podemos usar para nuestro estudio en **Redes de Regulación Genética**.

El primer análisis comparativo sistemático de métodos de detección de comunidades fue realizado por Danon, Diaz-Guilera, Duch y Arenas [265], quienes compararon los desempeños de varios algoritmos usando la *prueba Girvan-Newman* (Sección 3.4.2.2). Los algoritmos examinados se enumeran en la tabla 3.21, junto con su complejidad. La figura 3.22 muestra el rendimiento de todos los algoritmos. Danon *et al.* consideraron tres valores para $\langle k \rangle_{out} = z_{out}$ (6, 7 y 8), y representaron el resultado para cada algoritmo como un grupo de tres columnas (figura 3.22), indicando el promedio (de varias realizaciones) de la fracción de nodos correctamente clasificados (sección 3.4.3.1) entre la *partición plantada* y la *partición encontrada* por el método.

Author	Ref.	Label	Order
Eckmann & Moses	(Eckmann and Moses, 2002)	EM	$O(m\langle k^2\rangle)$
Zhou & Lipowsky	(Zhou and Lipowsky, 2004)	ZL	$O(n^3)$
Latapy & Pons	(Latapy and Pons, 2005)	LP	$O(n^3)$
Clauset et al.	(Clauset *et al.*, 2004)	NF	$O(n\log^2 n)$
Newman & Girvan	(Newman and Girvan, 2004)	NG	$O(nm^2)$
Girvan & Newman	(Girvan and Newman, 2002)	GN	$O(n^2m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	SA	parameter dependent
Duch & Arenas	(Duch and Arenas, 2005)	DA	$O(n^2\log n)$
Fortunato et al.	(Fortunato *et al.*, 2004)	FLM	$O(m^3n)$
Radicchi et al.	(Radicchi *et al.*, 2004)	RCCLP	$O(m^4/n^2)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM/DMN	$O(n^3)$
Bagrow & Bollt	(Bagrow and Bollt, 2005)	BB	$O(n^3)$
Capocci et al.	(Capocci *et al.*, 2005)	CSCC	$O(n^2)$
Wu & Huberman	(Wu and Huberman, 2004)	WH	$O(n+m)$
Palla et al.	(Palla *et al.*, 2005)	PK	$O(\exp(n))$
Reichardt & Bornholdt	(Reichardt and Bornholdt, 2004)	RB	parameter dependent

Figura 3.21: **Tabla comparativa de algoritmos en el estudio realizado por Danon *et al.***. Se muestra la complejidad algorítmica de cada método.

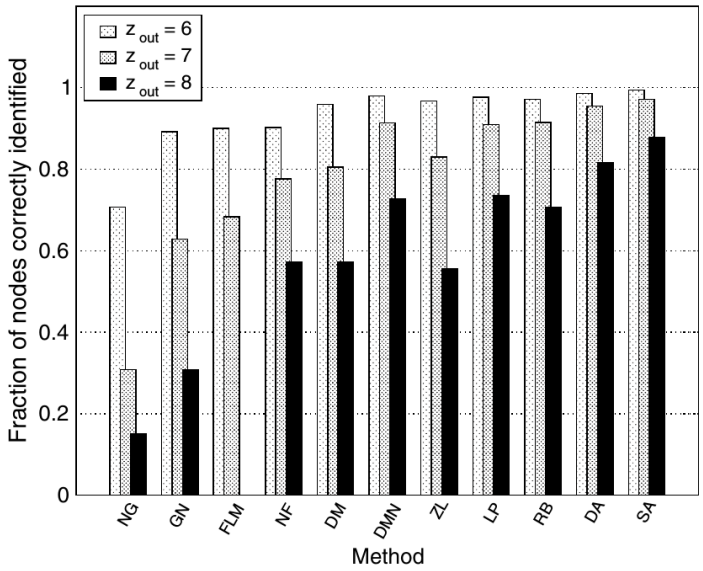


Figura 3.22: **Rendimiento de algoritmos en el estudio realizado por Danon *et al.***. Para cada método se tiene un grupo de tres columnas ($\langle k\rangle_{out} = z_{out}$ =6, 7 y 8), cuya altura representa la fracción de nodos correctamente clasificados de la *partición encontrada* respecto a la *partición plantada*.

La comparación muestra que el método de Guimerá *et al.* [145] de optimización de modularidad a través de templado simulado (sección 3.3.4.3) produce los mejores resultados, aunque es un procedimiento bastante lento, que no se puede aplicar a redes de tamaño del orden de 10^5 nodos o más. Además, como ya hemos señalado, la *benchmark Girvan-Newman* no es una buena representación de redes reales con estructura modular (que se caracterizan por distribuciones heterogéneas de grados y tamaños de comunidad). En este sentido, la clase de redes diseñados por Lancichinetti et al. (*benchmark LFR*) [262] (Sección 3.4.2.3) plantea una prueba mucho más severa para las técnicas de detección de comunudiades. Por ejemplo, muchos métodos tienen problemas para detectar comunidades de tamaños muy diferentes (como la mayoría de los métodos enumerados en la Tabla 3.21).

Dado lo anterior, Lancichinetti *et al.* han llevado a cabo un cuidadoso análisis comparativo de los métodos de detección de comunidades con el *benchmark LFR*. Los algoritmos elegidos se enumeran en la Tabla 3.23.

Author	Ref.	Label	Order
Girvan & Newman	(Girvan and Newman, 2002; Newman and Girvan, 2004)	GN	$O(nm^2)$
Clauset et al.	(Clauset *et al.*, 2004)	Clauset et al.	$O(n \log^2 n)$
Blondel et al.	(Blondel *et al.*, 2008)	Blondel et al.	$O(m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	Sim. Ann.	parameter dependent
Radicchi et al.	(Radicchi *et al.*, 2004)	Radicchi et al.	$O(m^4/n^2)$
Palla et al.	(Palla *et al.*, 2005)	Cfinder	$O(\exp(n))$
Van Dongen	(Dongen, 2000a)	MCL	$O(nk^2)$, $k < n$ parameter
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2007)	Infomod	parameter dependent
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2008)	Infomap	$O(m)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM	$O(n^3)$
Newman & Leicht	(Newman and Leicht, 2007)	EM	parameter dependent
Ronhovde & Nussinov	(Ronhovde and Nussinov, 2009)	RN	$O(m^\beta \log n)$, $\beta \sim 1.3$

Figura 3.23: **Tabla comparativa de algoritmos en el estudio realizado por Lancichinetti *et al.*, prueba LFR.**

En las figuras 3.24, 3.25 y 3.26 se muestran las comparaciones de los rendimientos de los métodos mediante el *benchmark LFR*, cada punto en la gráfica es un promedio de 100 realizaciones de la prueba. Las gráficas muestran la *Información Mutua Normalizada* (sec. 3.4.3.2) en función del *parámetro de mezcla* μ. Cada nodo comparte, una fracción $1-\mu$ de sus aristas con los nodos de su propia comunidad y una fracción μ con los nodos de las otras comunidades (ver sección 3.4.2.3). Por ejemplo si $\mu = 4$ quiere decir que los nodos tienen 60 % de sus conexiones dentro de su propia comunidad. La figura muestra diferentes curvas que refieren a diferentes tamaños de sistema (1000 y 5000 nodos) y rangos de tamaño de comunidad: (**S**) de 10 a 50 nodos y (**B**) de 20 a 100 nodos. Así entonces, si $0 \leq \mu < 0.5$ la estructura modular esta muy bien determinada. En su caso, se llevaron a cabo pruebas en las versiones del benchmark LFR con aristas dirigidas, pesadas y/o comunidades superpuestas [263, 264].

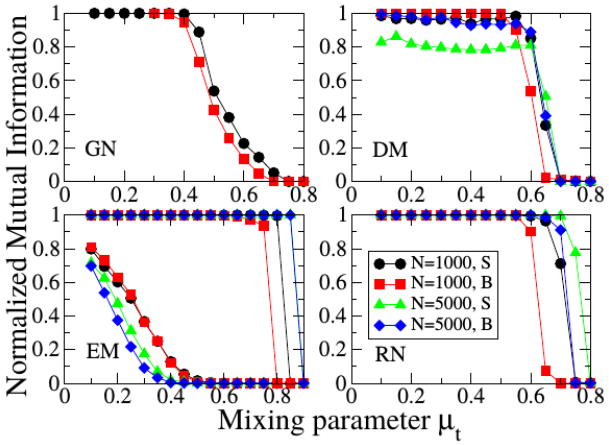


Figura 3.24: **Comparación de métodos bajo la *prueba LFR***. Se muestra el rendimiento del *algoritmo de Girvan-Newman* (sección 3.3.2.1) para el cual solo se adoptó el tamaño más pequeño, debido al alto tiempo de computo del método. Así también se muestra el rendimiento para el *algoritmo de Donetti y Muñoz* (sección 3.3.3.2) y para el *método de Ronhovde y Nussinov* (sección 3.3.5.1).

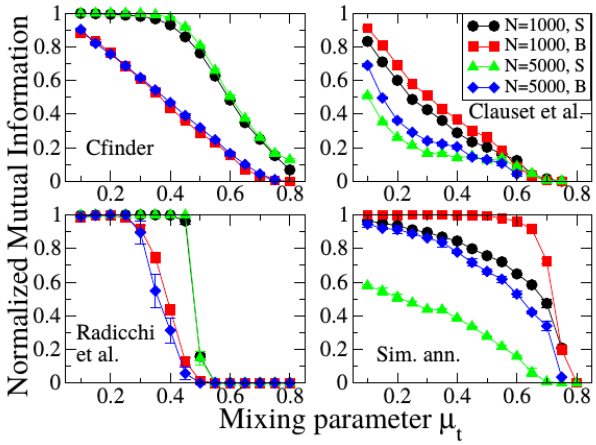


Figura 3.25: Se muestra el rendimiento del *Método de Percolación de Clique (Clique Percolation)* o *Cfinder* de Palla *et al.* (sección 3.3.6.3); el del *algoritmo de Radicchi et al.* (sección 3.3.2.2). Así como para el *algoritmo de Clauset-Newman* (sección 3.3.4.1) y el *Templado simulado (Simulated annealing)* de Guimerá *et al.* (sección 3.3.4.3).

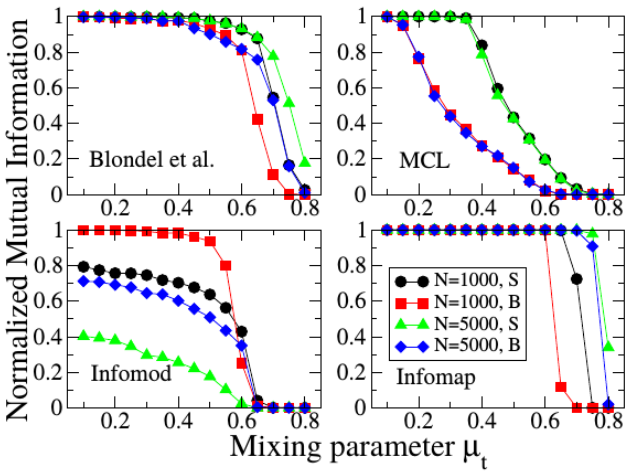


Figura 3.26: **Comparación de métodos bajo la *prueba LFR***. Se muestra el rendimiento del *método de Louvain* de Blondel *et al.* (sección 3.3.4.2); el del *Markov Cluster Algorithm* o MCL de Van Dongen (sección 3.3.5.4). Así como para **Infomap** de Rosvall y Bergstrom (sección 3.3.5.4).

El método **Infomap** de Rosvall y Bergstrom [156] (sección 3.3.5.4) muestra ser el mejor, pero también el *método de Louvain* de Blondel *et al.* [162] (sección 3.3.4.2) y el *método de Ronhovde y Nussinov* [244] (sección 3.3.5.1) tienen un buen desempeño. Estos tres métodos también son muy rápidos, con una complejidad que es esencialmente lineal con el tamaño n del sistema, por lo que se pueden aplicar a sistemas grandes. Por otro lado, los métodos basados en modularidad (con la excepción del *método de Louvain*) tienen un rendimiento bastante pobre, que empeora para sistemas más grandes y comunidades más pequeñas, debido al conocido límite de resolución de la modularidad (sección 3.3.4.5). El rendimiento de los métodos restantes empeora considerablemente si se aumenta el tamaño del sistema (como el de *Donetti y Muñoz*) o el tamaño de las comunidades (*e.g. Cfinder*, *Markov Cluster Algorithm* y *método de Radicchi et al.*).

Cabe señalar, que el *Método de Percolación de Clique (Clique Percolation)* o *Cfinder* de Palla *et al.* [175] (sección 3.3.6.3), diseñado para encontrar comunidades superpuestas y muy popular en aplicaciones a redes biológicas (sección 3.1.2) tiene un rendimiento pobre, ya que los nodos superpuestos encontrados por el algoritmo son en general diferentes de los nodos superpuestos de la *partición plantada* del *benchmark*. Asimismo, otro método que vale la pena mencionar por su bien desempeño en términos de la *prueba LFR*, incluso para redes muy grandes es el *Método de Optimización Local de Estadísticas de Orden* (*Order Statistics Local Optimization Method* u **OSLOM** de Lancichinetti *et. al.* [255] (sección 3.3.6.1). El cual ha mostrado un rendimiento comparable con **Infomap** de Rosvall y Bergstrom [156] (figura 3.27).

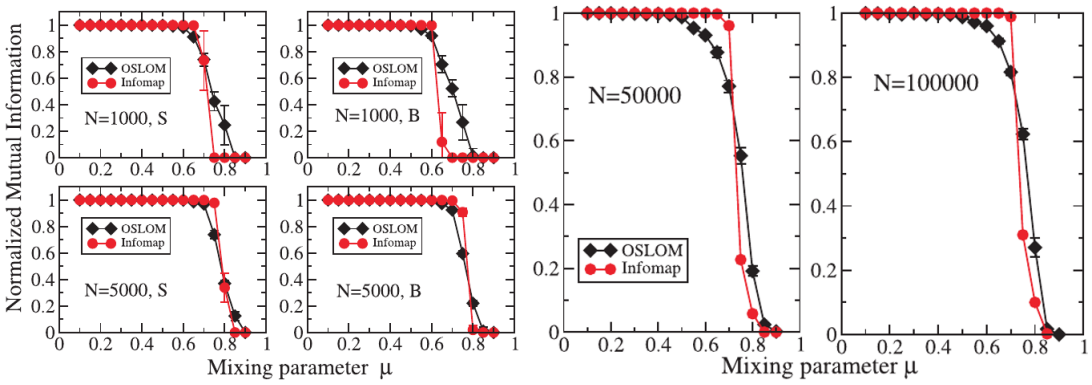


Figura 3.27: **Comparación de Infomap y OSLOM bajo la *prueba LFR***. El rendimeinto del *Método de Optimización Local de Estadísticas de Orden* (**OSLOM**) de Lancichinetti *et. al.* (sección 3.3.6.1) es comparable con el de **Infomap** de Rosvall y Bergstrom (sección 3.3.5.4). A la izquierda se muestra el rendimiento para diferentes tamaños de sistema correspondientes a la prueba LFR estándar (1000 y 5000 nodos) y rangos de tamaño de comunidad: (**S**) de 10 a 50 nodos y (**B**) de 20 a 100 nodos. A la derecha se muestra el rendimiento para redes muy grandes correspondientes a 50,000 y 100,000 nodos.

Así entonces, podemos definir algunos de los mejores métodos computacionales de detección de comunidades para las familias de técnicas expuestas. Para los métodos de **Optimización de Modularidad** (sección 3.3.4), *método de Louvain* de Blondel *et al.* [162] (sección 3.3.4.2) es uno de los mejores. En el caso de los **métodos espectrales** (sección 3.3.3), el *algoritmo de Donetti y Muñoz* [235] (sección 3.3.3.2 también es de lo más recomendables.

Asimismo, sin duda alguna los **Algoritmos Dinámicos** (sección 3.3.5) son los que muestran mejores rendimientos en general, entre estos podemos encontrar para las técnicas basadas en *Modelos de Spin* (sección 3.3.5.1) el *método de Ronhovde y Nussinov* [244] (sección 3.3.5.1) y para las técnicas basadas en *caminatas aleatorias* (sección 3.3.5.4) **Infomap** de Rosvall y Bergstrom [156] (sección 3.3.5.4) es la mejor opción dado que además se ha extendido a módulos jerárquicos y sobrepuestos. Finalmente **OSLOM** de Lancichinetti *et. al.* [255] (sección 3.3.6.1), también muestra muy buenos resultados.

Hay que agregar que dentro de estos métodos hay algunos cuya implementación (realizada por los mismos autores) tiene soporte constante y son fáciles de obtener y ejecutar. Estos son las casos del método de **Louvain**, el método de *Donetti y Muñoz*, **OSLOM** e **Infomap** aunque el el algoritmo de *Donetti y Muñoz* no es de los más

rápidos. Así, podemos ver que **Infomap** de Rosvall y Bergstrom [156] (sección 3.3.5.4) es una de las mejores propuestas tanto por eficiencia, exactitud y facilidad de aplicación.

Finalmente, es importante señalar que las ideas principales de este capitulo aplicadas al análisis de modularidad y detección de comunidades en redes biológicas se han condensado en un reciente articulo de revisión de Alcalá *et. al.* [266].

Capítulo 4

Detección de módulos biológicamente funcionales.

Una vez que se han expuesto a formalmente las principales características de las redes complejas y asimismo la relevancia de detectar su estructura modular, llevemos estás ideas a modelar el programa de regulación genético. En este capitulo expondremos nuestra metodología para asociar funciones a grupos de genes co-regulados a partir de datos experimentales reales, así como expondremos algunos casos de estudio donde hemos aplicado ésta metodología.

Como se abordó en los capítulos anteriores, las redes complejas tienen varias características topológicas que pueden ayudar a identificar procesos emergentes en sistemas complejos. Asimismo existen muchos métodos para detectar comunidades y varias consideraciones sobre la estructura modular de una red compleja. Usando estos formalismos teóricos es posible abordar las *Redes de Regulación Genética* de miles de genes con el objetivo de poder entender y vislumbrar algunos aspectos del programa de regulación genetica que dan paso a algunos fenotipos de interés en organismos con genomas grandes. Particularmente el caso de los seres humanos con la idea principal de identificar des-regulaciones que deriven en enfermedades.

Así entonces la propuesta es muy clara, identificar módulos en una red de regulación genética de miles de elementos e interacciones que representa la co-regulación concertada del *genoma* completo de un ser humano, podría ayudar a tener una descripción más fina de algunos aspectos de su programa regulatorio. Asimismo entender e identificar pequeñas partes de este programa que funcionen de manera autónoma podría ayudar a dirigir mejor los esfuerzos experimentales sobre el estudio molecular de la interacción genética en enfermedades. Así entonces identificar grupos de genes que estén corregulados entre si y que además estén asociados a una función biológica particular ayudaría a entender a un nivel más general una enfermedad como por ejemplo el cáncer. A continuación se expondrá la propuesta metodológica que proponemos para lograr este objetivo.

Propuesta Metodológica.

La Propuesta Metodológica que aquí presentamos para detectar *módulos biológicamente funcionales* en *redes de regulación genética* de muchos elementos, se puede resumir en tres aspectos muy generales: primero la inferencia de la red en estudio, luego la partición de la red en módulos utilizando el algoritmo *InfoMap* para después, realizar un análisis de enriquecimiento de las comunidades encontradas.

Las redes de regulación genética a las que podemos aplicar esta metodología en general son inferidas a partir de datos genómicos o transcriptomicos provenientes de experimentos de genoma completo o parcial. En ambos casos la condición es que sean redes con al menos un orden de miles de nodos (genes) y 10^4 aristas (interacciones transcripsionales). Aunque en general la hemos aplicado al caso humano (*homo sapiens*), también es posible usarla en datos provenientes de otros organismos como ratón (*mus musculus*), mosca de fruta (drosophila melanogaster), la planta *arabidopsis thaliana*, el gusano *Caenorhabditis elegans* y en general a cualquier organismo del que se tengan datos genómicos de los que sea posible inferir una red grande o bien sus interacciones transcripsionales hayan sido estudiadas y estén anotadas en bases de datos.

Una vez inferida la red, el siguiente paso es particionarla en módulos para lo cual hemos escogido el algoritmo *Infomap* (véase sección 3.3.5.4), un enfoque basado en teoría de la información que captura el flujo de la misma vía la descripción de una caminata aleatoria. Cómo veremos más adelante (sección 4.1.1) *Infomap* combina varias de las mejores ideas expuestas en el capitulo anterior, pero sobre todo es uno de los métodos más eficientes, precisos y confiables en términos de benchmarks como la *prueba LFR* (ver sección 3.4.2.3). Además de lo anterior, *Infomap* también es de los algoritmos más rápidos en términos de tiempos de ejecución y complejidad algorítmica, y cuenta con una excelente implementación proporcionada por los mismos autores.

El último paso, como veremos en la sección 4.1.2, es asociar los conjuntos de genes que se derivan de los módulos detectados en la red, a funciones biológicas concretas. Para esto se implementa un *Análisis de Enriquecimiento* realizando una *prueba hipergeométrica*, la cual provee un valor de *significancia* sobre si un conjunto genes está estadísticamente asociado a una función biológica anotada en alguna base de datos como *Gene Ontology*, *KEGG* o *PathwayCommons*.

Finalmente, una vez que se tienen identificados conjuntos de genes *corregulados transcripsionalmente* (módulos) y se han asociado a funciones biológicas particulares, es posible interpretar los resultados analizando estos grupos (pequeños respecto a todo el genoma) como unidades transcripsionales funcionales dentro del programa regulatorio. Como veremos en los diferentes casos de estudio que se presentarán (sección 4.2), analizar los módulos puede ayudar a vislumbrar asepectos de dichas sibunidades involucradas en alguna enfermedad. Asimismo pueden ser analizados como pequeñas sub-redes

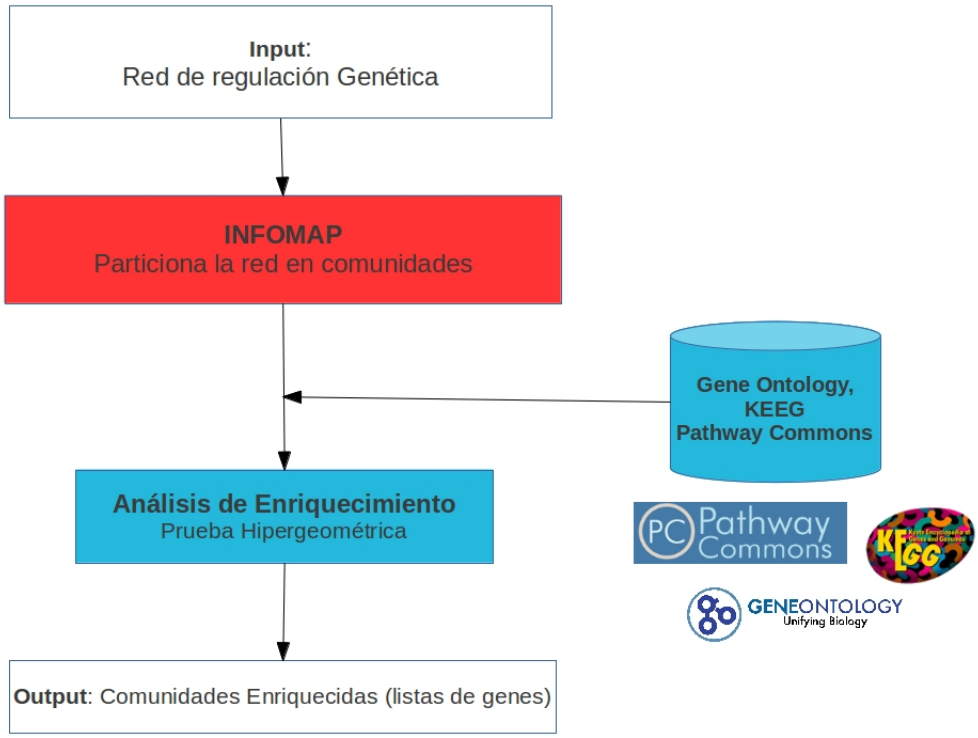


Figura 4.1: Propuesta Metodológica para la búsqueda de módulos funcionales en redes de regulación genética.

funcionales cuyos nodos tendrán diferentes centralidades en la red y es su mismo módulo; con lo que se puede buscar si tienen alguna relevancia reportada en la literatura, sobre algún fenotipo dada la función biológica a la que están asociados. Asimismo, si se cuentan con los *perfiles de expresión*[1] de dichos conjuntos de genes es posible comparar mediante un *Análisis de Expresión Diferencial* una posible des-regulación asociada a alguna función biológica de un fenotipo enfermo respecto del sano.

A continuación se detallarán los pasos para encontrar los módulos en una red mediante *Infomap* para posteriormente describir como es que es posible asociar los genes de dichos módulos a funciones biológicas anotadas en bases de datos mediante el *análisis de enriquecimiento*. Por último se detallará la inferencia de cada una de las redes utilizadas según el caso de estudio. Un resumen de esta metodología se muestra en la figura 4.1.

[1] nivel de producción de ARN mensajero o proteína de un gen

Detección de Comunidades e Infomap.

Como hemos expuesto, *Infomap* de Rosvall y Bergstrom [156] es por mucho una de las mejores propuestas para detectar módulos en redes complejas. A continuación se detallarán las ideas principales detrás de la *MapEquation* y el funcionamiento del algoritmo.

Como hemos mencionado *Infomap* es un método que se basa en la dinámica sobre la red más que en su estructura topológica. Así, la *MapEquation* [156] captura el flujo de información en la red, por lo que las comunidades consisten en grupos de nodos entre los cuales el flujo de información persiste por un largo tiempo. A grandes rasgos. Para describir el flujo de información en la red, se usa un caminata aleatoria para reportar dicho flujo (como *proxi*). Por *ergodicidad*, si dicha caminata es infinita, se llegara a una distribución estacionaria tanto para cuando se visitan *estructuras de la red* como para cuando se abandonan (dichas estructuras pueden ser nodos o módulos). Dichas distribuciones estacionarias se pueden codificar de forma optima en usando la *codificación de Huffman* [251] en una cadena binaria. Así entonces la caminata se puede describir en términos de la *teoría de la información* a partir de la longitud de la cadena binaría asociada a dicha caminata.

Si se tuviera una partición de la red en M en módulos, se podría tener una descripción de una caminata aleatoria entre los módulos (o estructuras más relevantes) en la red. Está caminata se puede codificar en una cadena binaría cuya longitud $L(M)$ sería mínima en el caso de tener la mejor partición de la red. Lo anterior se puede lograr mediante la ecuación del mapa (*MapEquation*). Así, la *MapEquation* aprovecha la dualidad entre encontrar la mejor estructura modular en una red y minimizar la longitud de la descripción de los movimientos de un caminante aleatorio en una red. Es decir, para cada partición modular dada de la red, existe un costo de información asociado a describir los movimientos del caminante aleatorio, o de flujo. Algunas particiones generan longitudes de descripción más cortas y algunas otras más largas. La partición con la *longitud de descripción* más corta es la que mejor captura la estructura modular de la red con respecto a la dinámica en la misma. A continuación expondremos lo anterior con más detalle.

Codificación de una caminata aleatoria infinita en la red.

Dada una red con n nodos (dirigida o no), la probabilidad $p_{i\to j}$ de que un caminante aleatorio pase del nodo i al nodo j para $i, j \in N$ viene dada por:

$$p_{i\to j} = \frac{1}{k_i} = \frac{1}{\sum_j A_{ij}} \tag{4.1}$$

Recordemos que si la red es dirigida la matriz de adyacencia es asimétrica y que $A_{ij} \neq A_{ji}$ por lo que puede ser que $A_{ij} = 1$ (hay un camino de i a j) pero que $A_{ji} = 0$

(no hay un camino de regreso de j a i). Asimismo, cuando los enlaces son pesados w_{ij} entonces la probabilidad condicional de que el caminante aleatorio pase del nodo i al nodo j está dado por el peso relativo del enlace:

$$p_{i \to j} = \frac{w_{ij}}{\sum_j W_{ij}} \tag{4.2}$$

Asumiendo el *limite ergódico* en una caminata aleatoria infinita, se tiene entonces una distribución de probabilidad estacionaria p_i para un nodo i, la cual puede derivarse en principio del siguiente sistema recursivo de ecuaciones:

$$p_i = \sum_j p_j(p_{j \to i}) \tag{4.3}$$

De está manera tenemos la distribución de probabilidad estacionaria p_i de que se visite un nodo i en la red o alguna estructura de la misma en una caminata aleatoria infinita. Está dinámica puede codificarse en una cadena binaría mediante la *codificación de Huffman* [251], la cual es la forma más optima de describir una distribución estacionaria de frecuencias en código binario. En dicha codificación entre más frecuente es un variable aleatoria X se le asocia una *palabra código* (*codeword*) más corta. De está manera, si es más probable visitar un nodo, o una estructura (módulo) en la red, esta tendrá asociada una *codeword* más corta.

Cuando la red es dirigida existe la posibilidad de que el caminante aleatorio quede atrapado en un nodo que no tenga conexiones salientes. Así entonces, para garantizar una distribución de estado estacionaria única independiente de donde comience el caminante aleatorio en redes dirigidas, la probabilidad se corrige con una pequeña probabilidad de *teletransportación* τ en la caminata aleatoria, que vincula a todos los nodos entre si con una probabilidad positiva, y por lo tanto, el caminante aleatorio se convierte en un *"surfer"* aleatorio. Esta misma estrategia ayuda cuando se tienen redes disconexas, es decir que se componen de varias "*islas*" (componentes). Está corrección vía el parámetro τ, permite que el caminante no alcance la distribución estacionaria considerando sólo en el componente más grande. De esta manera, la distribución estacionaria corregida con el parametro τ está dada por:

$$p_i = (1 - \tau) \sum_j p_j(p_{j \to i}) + \tau \frac{\sum_j w_{ji}}{\sum_{i,j} W_{ji}} \tag{4.4}$$

Lo anterior aún no nos dice cómo particionar la red en módulos, pero si nos dice como codificar una caminata aleatoria entre nodos o *estructuras de la red* de acuerdo a la probabilidad de ser visitadas. Asimismo sabemos que en términos de la codificación de Huffman una cadena corta corresponde a estructuras con frecuencias altas de vista por un caminante aleatorio en la red. Así entonces si pudiéramos tener una descripción corta de una caminata en la red, entonces tendríamos una descripción de las estructuras más relevantes en la misma, tal y como hacen los mapas cuando describen grandes regiones

de una ciudad o un país mediante una palabra corta en vez de una descripción larga (por ejemplo, una dirección) de un sitio muy particular. La pregunta ahora es entonces: ¿cómo encontrar la codificación optima (de *longitud* más corta) que describa los módulos de la red en términos de una caminata aleatoria infinita en la misma? Esta pregunta se puede responder usando *teoría de la información*.

Teoría de la información y la *MapEquation*

El teorema de codificación de Shannon [61, 62, 63, 64, 65] establece que el *límite inferior teórico* para describir un flujo de n variables aleatorias independientes e idénticamente distribuidas viene dado por la entropía de la distribución de probabilidad. Es decir, dada la distribución de probabilidad $\mathcal{P} = \{p_i\}$ tal que $\sum_i p_i = 1$, el límite inferior de la *longitud de código* está dado por:

$$L(\mathcal{P}) = H(\mathcal{P}) \equiv -\sum_i p_i log(p_i) \tag{4.5}$$

con el logaritmo tomado en base 2 para medir la *longitud de código* en bits. En consecuencia, la mejor compresión de la dinámica del caminante aleatorio en una red viene dada por:

$$\sum_i p_i H(p_{i\to j}) \tag{4.6}$$

Esta tasa de entropía [61, 62] corresponde a la *longitud de código* promedio para especificar la siguiente visita a un nodo partiendo del nodo actual, promediada sobre todos los nodos. Este esquema de codificación aprovecha las visitas a los nodos siguientes, pero no aprovecha la estructura modular de la red. Por lo que la *mapequation* usa la restricción adicional de que la caminata aleatoria también ocurre entre módulos. A partir de esta suposición, es posible lograr una descripción (*codificación*) modular (comprimida al máximo) a partir de la partición que mejor represente la estructura modular de la red con respecto a la dinámica en la misma.

Así, suponiendo que se tiene una partición M de red, para n nodos en s módulos, con cada nodo i asignado a un módulo c_α ($\alpha \in 1, \ldots, s$), la *mapequation* especifica la *longitud teórica* de la descripción modular de la trayectoria de un caminante aleatorio, guiada por los enlaces (en su caso dirigidos y pesados) de la red. De ésta manera, la *mapequation* se puede expresar invocando el teorema de codificación de Shannon [61, 62] (ecuación 4.5) a partir de las tasas de visita p_i a los nodos y las tasas de transición $q_{c_\alpha \curvearrowleft}$ y $q_{c_\alpha \curvearrowright}$ con las que el caminante entra y sale de cada módulo c_α, respectivamente. Dichas probabilidades de entrar a un módulo c_α (desde un módulo c_β) $q_{c_\alpha \curvearrowleft}$ y de salir $q_{c_\alpha \curvearrowright}$ hacia un módulo c_β (donde $\alpha \neq \beta$) están dadas por:

$$q_{c_\alpha \curvearrowleft} = \sum_{i \in c_\beta, j \in c_\alpha} q_{i\to j} \tag{4.7}$$

$$q_{c_\alpha \curvearrowright} = \sum_{i \in c_\alpha, j \in c_\beta} q_{i \to j} \tag{4.8}$$

Donde $q_{i \to j}$ es la probabilidad de ir a un nodo i desde el nodo j ($i \neq j$). Con lo que podemos calcular la probabilidad total de que el caminante aleatorio cambie de módulo:

$$q_{\curvearrowleft} = \sum_{\alpha=1}^{s} q_{c_\alpha \curvearrowleft} \tag{4.9}$$

Así entonces, denotamos como $\mathcal{Q}$ la distribución de probabilidad normalizada de de movimientos entre módulos y como $H(\mathcal{Q})$ a su entropía que es el límite inferior teórico de la longitud promedio de una *palabra código* (*codeword*) utilizada para nombrar un módulo.

$$H(\mathcal{Q}) = -\sum_{\alpha=1}^{s} \left(\frac{q_{c_\alpha \curvearrowleft}}{q_{\curvearrowleft}} \right) log \left(\frac{q_{c_\alpha \curvearrowleft}}{q_{\curvearrowleft}} \right) \tag{4.10}$$

Asimismo, para calcular la entropía de los movimientos dentro de un módulo c_α hay que calcular la probabilidad total de que el caminante aleatorio salga del módulo (y se use la *palabra código* de salida) más la probabilidad de que se visite cualquier nodo dentro del módulo:

$$p_{\circlearrowright}^{\alpha} = q_{c_\alpha \curvearrowright} + \sum_{i \in c_\alpha} p_i \tag{4.11}$$

De esta manera, denotamos como $\mathcal{P}_\alpha$ a la distribución de probabilidad normalizada de los movimientos dentro del módulo c_α (cuando el caminante visita cada nodo en el módulo c_α y/o sale del mismo) y como $H(\mathcal{P}_\alpha)$ a su entropía que es el lmite inferior teórico de la longitud promedio de una *palabra código* (*codeword*) utilizada utilizada para nombrar un nodo (incluido el código de salida) en el módulo c_α. De tal manera que:

$$H(\mathcal{P}_\alpha) = -\left(\frac{q_{c_\alpha \curvearrowright}}{p_{\circlearrowright}^{\alpha}} \right) log \left(\frac{q_{c_\alpha \curvearrowright}}{p_{\circlearrowright}^{\alpha}} \right) - \sum_{i \in \alpha} \left(\frac{p_i}{p_{\circlearrowright}^{\alpha}} \right) log \left(\frac{p_i}{p_{\circlearrowright}^{\alpha}} \right) \tag{4.12}$$

En resumen, para aprovechar la estructura modular de la red, hay que considerar la entropía asociada a las *palabras código* para describir los movimientos del caminante aleatorio dentro y entre los módulos, respectivamente. Las longitudes de las *palabras código* para las entradas a un módulo, se derivan del conjunto de frecuencias $q_{c_\alpha \curvearrowleft}$ con las cuales el caminante ingresa a cada módulo, donde $q_{\curvearrowleft}$ denota la suma de estas frecuencias, es decir, el uso total de las *palabras código* para moverse entre módulos, y $\mathcal{Q}$ denota la distribución de probabilidad normalizada de dichos movimientos. Asimismo

las longitudes de las *palabras código* de los movimientos del caminante aleatorio dentro de los módulos, se derivan de las frecuencias con las que el caminante aleatorio visita cada uno de los nodos en el módulo, p_i ($i \in c_\alpha$), y con la que sale del módulo, $q_{c_\alpha \curvearrowright}$. Asimismo $p^\alpha_{\circlearrowright}$ denota la suma de estas frecuencias, es decir, el uso total de *palabras código* para moverse dentro del módulo c_α; y finalmente $\mathcal{P}_\alpha$ denota la distribución de probabilidad normalizada de los movimientos dentro del módulo c_α.

Así entonces, la *Ecuación de Mapa* (*mapequation*) es:

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{\alpha=1}^{s} p^\alpha_{\circlearrowright} H(\mathcal{P}_\alpha) \tag{4.13}$$

Donde el primer término (en rojo) da la cantidad promedio de bits necesaria para describir el movimiento entre módulos, y el segundo término (en azul) da el número promedio de bits necesarios para describir el movimiento dentro de los módulos. Así entonces, se tiene la *descripción* de una caminata aleatoria entre estructuras de la red (módulos y/o nodos) con longitud L dada una partición M de la red (en s módulos).

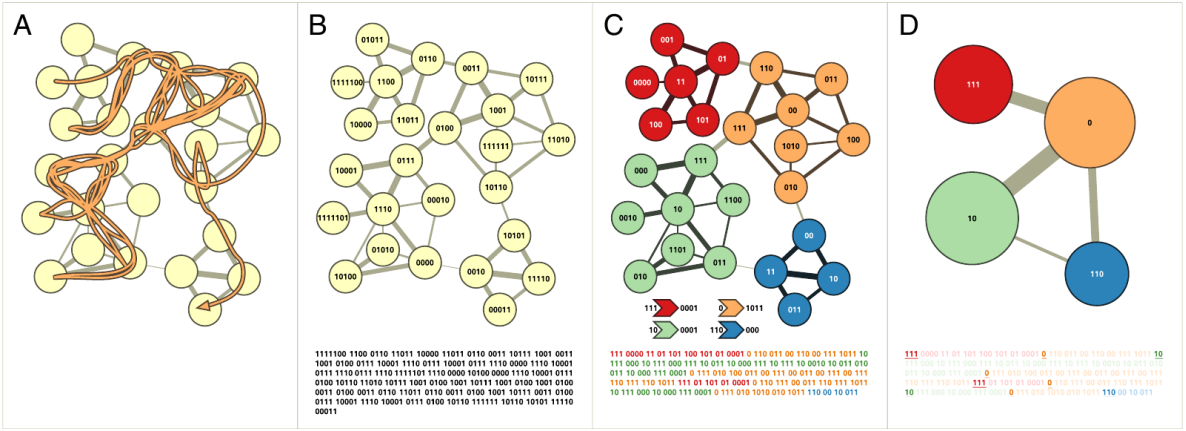


Figura 4.2: **Infomap y la *MapEquation***. **A**: Una caminata aleatoria alcanza una distribución estacionaria de visitas a los nodos en la Red. **B**: La dinámica de esta caminata puede codificarse la *codificación de Huffman*. **C**: Usando una descripción de la caminata más eficiente se pueden localizar estructuras (módulos y/o nodos) con mayor información en la red. **D**: la *mapequation* describe el movimiento entre módulos y dentro de los mismos, mediante la descripción optima de la caminata, lo que descubre los módulos de la red; nótese las transiciones entre los módulos están resaltadas en la *codificación* (parte inferior).

De esta manera, el problema de encontrar estructuras modulares (comunidades) en la red, se ha convertido en un problema de optimización cuyo objetivo es minimizar la función $L(M)$ que da la descripción optima de la caminata. Para lograrlo a partir de la *mapequation* hay que balancear los términos en rojo y azul (los movimientos entre módulos con los movimientos dentro de un módulo), para calcular la longitud mínima de descripción de una caminata aleatoria en la red, que separe las estructuras importantes de los detalles no significativos y así obtener la mejor partición M de la red en comunidades. Como veremos a continuación, para lograr este objetivo, *Infomap* se basa fuertemente en el *método de Louvain* [162] (ver sección 3.3.4.2) además de utilizar otras técnicas de optimización para refinar el resultado.

Minimización de $L(M)$ y refinamiento por *Templado Simulado*

La elección de una partición M debe reflejar los patrones de flujo dentro de la red, con cada módulo correspondiente a un grupo de nodos en el que el caminante aleatorio pase un largo período de tiempo antes de partir hacia otro módulo. Para encontrar la mejor partición de este tipo, hay que minimizar la *ecuación de mapa* sobre todas las particiones posibles M, la partición que proporcione la *longitud de descripción* más corta captará mejor la estructura de la red con respecto a la dinámica en la misma.

Así entonces, el algoritmo procede de manera muy similar al enfoque *greedy* del *método de Louvain* propuesto por Blondel *et. al.* [162] (ver sección 3.3.4.2). Primero los nodos vecinos se unen en módulos, que posteriormente se unen en supermódulos y así sucesivamente. Al principio, cada nodo está asignado a su propio módulo. Luego, en un orden secuencial aleatorio, cada nodo se mueve al módulo vecino que resulte en la mayor disminución de la *mapequation* (ecuación 4.13). Si ningún movimiento resulta en una disminución de la *mapequation*, el nodo permanece en su módulo original. Este procedimiento se repite, cada vez en un nuevo orden secuencial aleatorio, hasta que ningún movimiento genera una disminución de la *ecuación de mapa*. Luego se reconstruye la red, con los módulos del último nivel formando nuevos nodos en este nivel y, exactamente como en el paso anterior los nodos se van uniendo en nuevos supermódulos.

Esta reconstrucción jerárquica de la red se repite hasta que la *mapequation* no se pueda reducir más (es decir, hasta que la *longitud de descripción* no se pueda minimizar más). Con este procedimiento, se puede encontrar una agrupación bastante buena de la red en muy poco tiempo. Sin embargo se puede refinar aín más usando la técnica de *Templado Simulado* (*Simulated annealing*) [240] al estilo propuesto por R. Guimerá *et al.* [145] (ver sección 3.3.4.3).

Cuando la red se reconstruye, dos o más módulos que se fusionan y forman un único módulo no se pueden volver a separar. Así nodos asignados al mismo módulo se ven obligados a "moverse" conjuntamente durante el resto del procedimiento, por lo que puede ocurrir que lo que fue un movimiento óptimo al principio del algoritmo podría tener

el efecto opuesto más adelante. Por lo tanto, la precisión puede mejorarse rompiendo módulos en el estado final en cualquiera de las dos formas siguientes:

Movimientos de submódulos (*movimientos globales*). Cada módulo se trata como una red en sí misma y el algoritmo principal se aplica a estas sub-redes. Este procedimiento genera uno o más submódulos por cada módulo. Luego todos los submódulos (como si fueran nodos) se mueven de vuelta a sus respectivos módulos del paso anterior. Y entonces, bajo la misma partición que en el paso anterior pero con cada submódulo libre de moverse entre módulos, el algoritmo principal se vuelve a aplicar.

Movimientos de un solo nodo (*movimientos locales*). Cada nodo se vuelve a asignar como el único miembro de su propio módulo. Luego, todos los nodos vuelven a sus respectivos módulos del paso anterior. Y entonces, bajo la misma partición que en el paso anterior pero ahora con cada nodo individual libre de moverse entre módulos, se vuelve a aplicar el algoritmo principal.

En la práctica, se repiten estás dos extensiones al algoritmo central en secuencia, siempre y cuando se mejore la partición (es decir, se minimice $L(M)$). Además los *movimientos de submódulo* se hacen recursivamente, es decir, para encontrar los submódulos que se moverán, el algoritmo primero divide los submódulos en subsubmódulos y subsubsubmódulos, y así hasta que no se puedan dividir más. Finalmente, dado que el algoritmo es estocástico y rápido, se puede reiniciar desde cero cada vez que la partición no pueda mejorarse más y el algoritmo se detenga. Para cada iteración, se registra si la *longitud de la descripción* $L(M)$ es más corta que la longitud registrada anteriormente. Y dado que la implementación es sencilla, si se repite la búsqueda más de una vez (100 veces o más si es posible), es menos probable que la partición final corresponda a un mínimo local.

Detección de Submodularidad Jerárquica.

Por ultimo, es importante mencionar que existe una versión de *Infomap* para encontrar la estructura modular jerárquica en redes, a partir de una generalización de la *mapequation* [253], la cual es parte de los resultados mostrados en el presente trabajo.

En general, la *mapequation* jerárquica (ecuación 4.14) libera la restricción de tener un solo índice para describir los movimientos dentro de los módulos y permite un número arbitrario índices anidados jerárquicamente que especifican los movimientos entre módulos, submódulos, subsubmódulos, y así sucesivamente, hasta el nivel más fino.

$$L(M) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{\alpha=1}^{s} L(M^{\alpha}) \tag{4.14}$$

Donde con la *longitud de descripción* del *sub-mapa* M^{α} en niveles intermedios está dada

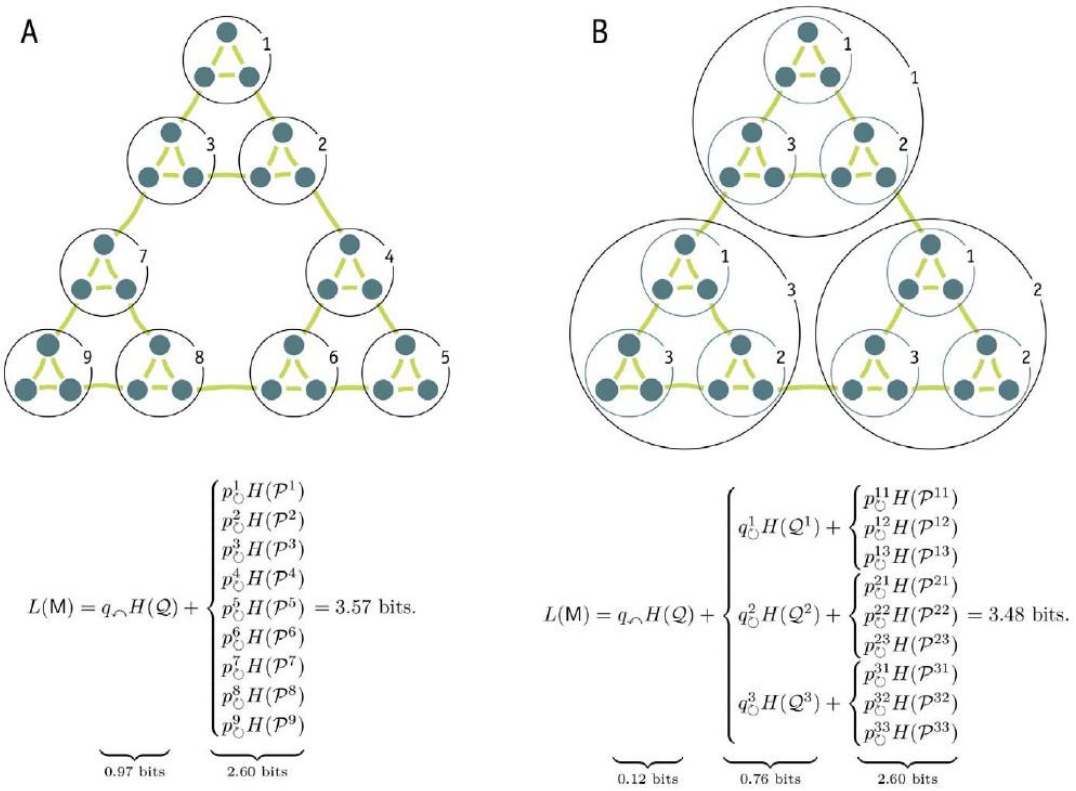


Figura 4.3: Ejemplo del uso de la *MapEquation* Jerárquica.

por:

$$L(M^{\alpha}) = q^{\alpha}_{\circlearrowright} H(\mathcal{Q}^{\alpha}) + \sum_{\beta=1}^{s_{\alpha}} L(M^{\alpha\beta}) \tag{4.15}$$

y en el nivel modular más fino por:

$$L(M^{\alpha\beta\ldots\gamma}) = p^{\alpha\beta\ldots\gamma}_{\circlearrowright} H(\mathcal{P}_{\alpha\beta\ldots\gamma}) \tag{4.16}$$

Está búsqueda recursiva opera en un módulo a cualquier nivel; esto puede ser en todos los nodos en toda la red, o en algunos nodos en el nivel más fino. Para un módulo dado, el algoritmo primero genera submódulos si esto le da una *longitud de descripción* $L(M)$ más corta. Si no, la búsqueda recursiva no va más allá en esta rama. Pero si al agregar submódulos se obtiene una longitud de descripción más corta, el algoritmo prueba si los movimientos dentro del módulo se pueden comprimir aún más.

Se puede lograr una compresión adicional añadiendo calculando la entropía de los movimientos dentro de los submódulos. Para probar todas las combinaciones, el algoritmo se llama recursivamente, tanto operando en la red formada por los submódulos como en las redes formadas por los nodos dentro de cada submódulo. De esta forma, el

algoritmo aumenta y disminuye sucesivamente la profundidad de las diferentes ramas de la estructura multinivel en su búsqueda de la partición jerárquica óptima. Por cada división de un módulo en submódulos, se el algoritmo descrito anteriormente.

Análisis de Enriquecimiento estadístico.

Una vez que se han obtenido los módulos de la red mediante *Infomap* (sección 4.1.1), el paso siguiente es verificar si las listas de genes provenientes de dichos módulos, que a su vez representen sub-unidades de interacciones en la red (sub-unidades en el modelo de corregulación), están asociadas a alguna función biológica particular, o dicho más formalmente, si a alguna función biológica está *sobre-representada* en los módulos. Lo anterior se logra realizando un *Análisis de Enriquecimiento estadístico* sobre categorías de genes cuya función es conocida o está anotada en alguna base de datos. Para esto, se pueden usar bases de datos como las contenidas en la *Kyoto Encyclopedia of Genes and Genomes* (**KEEG**), *PathwayCommons* (**PC**) o el *Gene Ontology Consortium*[184] (**GO**).

Figura 4.4: **Base de Datos usadas para análisis de enriquecimiento**. Las bases de datos como *Kyoto Encyclopedia of Genes and Genomes* (**KEEG**), *PathwayCommons* (**PC**) o el *Gene Ontology Consortium* (**GO**), tienen anotadas funciones biológicas bien estudiadas para genes.

El llamado *análisis de enriquecimiento* [267, 268, 269], es un enfoque que ha demostrado ser bastante útil para lograr una descripción bien fundada de procesos funcionales, sobre células fenotípicas en estudios moleculares a gran escala de sistemas biológicos [270] y se basa en determinar la *sobrerrepresentación estadística* de genes sobre funciones biológicas anotadas en las bases de datos. Hay varias maneras de realizar análisis de enriquecimiento, pero la base lógica en general es la misma: si hay un número relativamente grande de moléculas asociadas con un proceso biológico particular en el sistema, existe una probabilidad de que dicho proceso este activo. Para especificar qué es un número "relativamente grande", necesitamos saber cuántas moléculas están pre-

sentes con respecto a las que son representativas de los procesos y comparar esto con un *modelo nulo* para evaluar la *significación estadística*. La forma más común de hacerlo es mediante un ***análisis de sobre-representación*** implementando una *prueba hipergeométrica* (o modelo de urna) generalmente corregida.

Análisis de Sobre-Representación y Prueba Hipergeométrica.

La hipótesis básica en un análisis de sobre-representación (*Over-Representation Analysis* - ORA) es que se pueden detectar funciones biológicas relevantes si la proporción de genes anotados dentro de alguna de estas funciones, excede la proporción de genes que podrían esperarse aleatoriamente. Así, mediante una prueba hipergeométrica[1] es posible medir la *significancia estadística* de que un conjunto de genes (un módulo) esté asociado a alguna función o proceso biológico (categoría).

Dicha *significancia estadística* se obtiene a partir de calcular la probabilidad de que un conjunto de igual tamaño de genes elegidos aleatoriamente resulten anotados en dicha función o proceso biológico (*hipótesis nula*). Para esto se usa un *modelo de urna* como modelo nulo. Se conoce como modelo de urna porque es equivale al siguiente experimento: Supongamos que los genes son canicas y que en una urna hay N canicas (el genoma completo). Se sabe que K de las N canicas son rojas (las de la categoría). Ahora bien si sacamos n canicas de la urna *al azar*, ¿cual es la probabilidad de que k de estás sean rojas? Esta probabilidad viene dada mediante la *distribución hipergeométrica*:

$$P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{4.17}$$

donde, N es el tamaño del genoma (o conjunto total de genes), K es el número de genes anotados en la función biológica en la base de datos, n es el tamaño de la lista de genes a probar (para este caso, cada módulo) y k es el número de genes que resulten en la categoría (éxitos).

En otras palabras $P(k)$ es la probabilidad de que aleatoriamente k de los n genes del módulo, estén asociados a una función biológica, sabiendo que hay K genes anotados en la misma, de los N de todo el genoma completo.

Así, la probabilidad de que un solo gen (del módulo) pertenezca a la categoría al azar ($k = 1$) es relativamente alta, mientras que la probabilidad de que todos los genes

[1] La prueba hipergeométrica es equivalente a la versión de una cola de la *Prueba Exacta de Fisher* [271].

(del módulo) pertenezcan a la categoria al azar ($k = n$) es muy, muy baja. Por lo que en el *análisis de sobre-representación* (para una función biológica conformada por K genes) se niega *esta hipótesis nula*, es decir, se mide la *signficancia estadística* de que un módulo este asociado a una función biológica imponiendo la condición que los genes (en dicho módulo) no pertenezcan a la categoría por azar. Lo anterior se mide mediante un *score de signficancia estadística* o *p-value* (p_v). Dicho *p-value* se calcula como la probabilidad de que se cumpla la hipótesis nula, es decir que aletaoriamnte k o más genes (del módulo) pertenezcan a la categoría.

$$p_v \equiv P(k) \tag{4.18}$$

Lo anterior implica que entre menor sea el p_v menor es la probabilidad de que el conjunto de n genes pertenezcan **al azar** a la categoría; y por lo tanto representará una *confianza estadística* sobre el conjunto de genes que estamos probando. Es decir, entre menor sea el p_v tenemos mayor *confianza estadística* de que los genes (en el módulo) estén participando de manera conjunta y no de manera azaroza en el mismo proceso o función biológica. Así, si fijamos el *p-value* (p_v), estamos garantizando un umbral de confianza estadística, es decir que la categoría (función biológica) este *sobre-representada* por más genes en un módulo dado. Para los casos de estudio a los que aplicamos esta metodología (sección 4.2) usamos umbrales de $p_v < 10^{-3}$ y $p_v < 10^{-5}$, que son bastante astringentes, pero como veremos en los resultados obtenidos, muchas de las categorías encontradas están *sobre-representadas* en módulos con *p-values* de hasta $p_v < 10^{-20}$ o menos. Además, para corregir posibles falsos positivos, es decir, posibles errores muestreos de pruebas múltiples a partir de $P(k)$, usamos la corrección del *algoritmo Benjamini-Hochberg*, denominada Tasa de Falsos Descubrimientos (*False Discovery Rare*) o FDR [272]. Solo asociaciones cuyos *p-values* corregidos oscilaron por debajo de los umbrales mencionados se consideraron significativos.

De esta manera, si un proceso o función biológica está *sobre-representada* en algún módulo de la red, decimos que dicha función biológica está ***enriquecida*** en ese módulo. Por lo que podemos asociar *estadísticamente* la función biológica al módulo de la red, con la confianza estadística que nos proporcione el *p-value* usado.

Lo anterior implica por lo tanto, que los módulos que estamos asociando a una funciones biológicas distan mucho de estar elegidos al azar, sino que cuentan con una confianza estadística dictada por un p_v. Esto es sumamente importante, pues significa que bajo nuestro modelo de red compleja, estamos localizando grupos de genes (dentro de todo un genoma) a partir de las interacciones en la red, es decir, a partir de que estos genes estén corregulados de una manera coordinada y que además participan de forma conjunta en proceso biológicos particulares conocidos. Por lo que podemos inferir *sub-unidades regulatorias funcionales* en fenotipos a partir de la estrcutra de se red de regulación, lo cual puede dirigir futuros estudios experimentales.

Análisis de Expresión Diferencial.

El *nivel de expresión* de un gen es una medida de cuánto se encuentra presente ("*expresado*") en la célula de algún fenotipo dado. Esto puede ser la medida de la cantidad de ARN mensajero (*transcrito*) presente en la célula o bien la cantidad proteína asociada al gen.

Cuando se cuentan con datos provenientes de muestras de genoma completo de humano, asociados a enfermedades (*e.g.* de cáncer), es posible identificar al menos dos fenotipos: enfermo y sano. Es decir que se sabe de antemano si la muestra exhibe el alguno de los dos fenotipos. Así, si unos de los datos es el *nivel de expresión génica* de cada muestra, es posible comparar las diferencias entre estos fenotipos: enfermo respecto del sano. Está comaparación se conoce como *Análisis de Expresión Diferencial*, y es útil para comparar a nivel genético y genómico muestras con diferentes fenotipos.

Para de los casos de estudio referentes a *Cáncer de Mama* a los que aplicamos nuestra metodología (sección 4.2), se contaba con el nivel de expresión de las diferentes muestras. Por lo que fue posible realizar un *Análisis de Expresión Diferencial* para complementar la metodología. Para esto, se normalizaron los niveles de expresión de las muestras sobre todo el genoma, para contar con un Nivel de Expresión Normalizado (NGE).

Casos de Estudio.

Ahora que se ha expuesto nuestra metodología para detectar *módulos funcionales* en rede de regulación genética presentaremos algunos casos de estudios a los que hemos aplicado dicha metodología.

Red transcripcional de MEF2C.

En este caso de estudio, consideramos el caso particular del gen *MEF2C*, que es un *factor de transcripción* (TF) asociado con el crecimiento muscular y procesos de desarrollo [273], el cual codifica una proteína de 473 aminoácidos (figura 4.5). Además, es un miembro de la familia *Mef2* que controla la expresión génica y es capaz de regular la diferenciación y el desarrollo celular [274]. Las moléculas MEF2 son reguladores altamente versátiles que comúnmente actúan como activadores de factores de transcripción. Sus interacciones proteína-proteína, principalmente con otros factores de transcripción, potenciadores o reguladores epigenómicos, junto con su actividad transcripcional de sitio de unión inherente, hacen que MEF2C sea un *regulador maestro (MR)* funcional y adaptable [275].

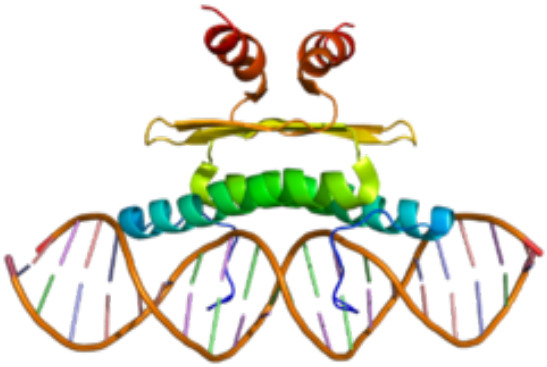


Figura 4.5: Factor de Transcripción MEF2C.

Se infirió una red de *sitios de enlace de factor de transcripción* (*transcription factor binding sites*, TFBS), es decir, la red que consiste en todos los *blancos transcripcionales* descendentes (hasta tres capas) de la molécula MEF2C (figura 4.6), algunos de estos *blancos* pueden incluso ser factores de transcripción de algunas moléculas en capas hacia arriba, que dan como resultado una topología de red compleja.

Usando la metodología propuesta (sección 4.1), se detectó con *InfoMap* (sección 4.1.1) la estructura modular en la red de regulación transcripcional. Una vez que se detectados los módulos (comunidades), se analizó si están involucradas en procesos biológicos específicos.

Construcción de la red a partir de Análisis de sitio de unión de *Factor de Transcripción* (TFBS).

Para construir la red de regulación transcripcional de MEF2C, se consideraron todos los *blancos transcripcionales* de MEF2C (figura 4.6) anotados en la base de datos **FANTOM4**[1] cuyo procedimiento experimental se detalla a continuación.

FANTOM4 infiere si un gen X es *Factor de Transcripción* de otro gen Y, si el promotor asociado con un *motivo TSS* en Y tiene un *motivo TFBS* regulatorio predicho en X, y dicha interacción tiene soporte experimental independiente [211]. Esto se logra usando dos métodos. En primer lugar usa técnicas de secuenciación profunda (***deepCAGE***) para supervisar *sitios de inicio de transcripción* (*motivos TSS*) en un nivel de resolución de par de base único, con lo que es posible identificar promotores activos y definir regiones relevantes para llevar a cabo predicciones de *sitios de unión de Factor de Transcripción* (TFBS). Por otro lado, para inferir patrones de TFBS conservados en ADN, se usa el algoritmo ***MotEvo*** [276], el cual es un enfoque Bayesiano

[1] `http://fantom.gsc.riken.jp/4/edgeexpress/view`

para integrar múltiples alineamientos de secuencia de ADN usados para inferir interacciones de regulación transcripcional. El soporte experimental para estas técnicas de alto rendimiento se realizó utilizando bases de datos como **JASPAR**[1] o **TRANSFAC**[2] que proporciona un conjunto de perfiles curados y no redundantes, derivados de colecciones publicadas de sitios de unión de factores de transcripción definidos experimentalmente para eucariotas, sitios de unión demostrados experimentalmente, secuencias de unión de consenso (matrices de peso posicionales) y genes regulados reportados.

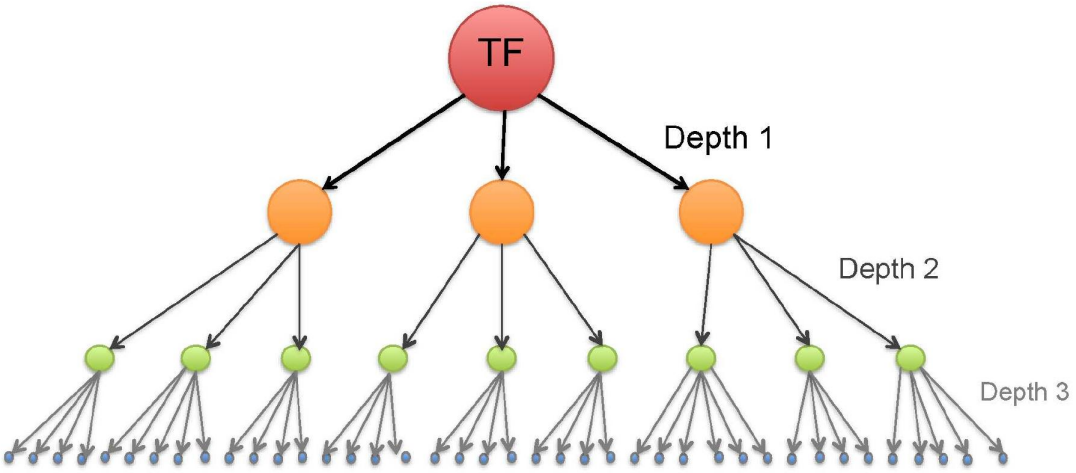


Figura 4.6: Red a de blancos transcripcionales de MEF2C a tres niveles de profundidad.

Dado lo anterior, FANTOM4 es una fuente confiable de información para construir una red de regulación génica basada en factores de transcripción. Así entonces, usamos la información anotada en esta base datos, para construir la red transcripcional de MEF2C hasta en tres niveles, es decir, los blancos transcripcionales de MEF2C (primer nivel); los blancos de esos blancos (segundo nivel) y los blancos de estos últimos (tercer nivel). Vale la pena mencionar que algunos de los blancos de primer, segundo y tercer nivel pueden ser a su vez reguladores de los niveles superiores, lo que indica la presencia de ciclos en la red y loe confiere una topología de red compleja (muy *sparse*) diferente a una red arbol.

Análisis de Enriquecimiento funcional.

Para el análisis de *enriquecimiento estadístico*, en este caso se realizó mapeando rutas (*Pathways*) biológicamente funcionales comunes para un determinado conjunto de

[1] `http://jaspar.genereg.net/`

[2] `http: //www.gene-regulation.com/pub/databases.html`

genes. Para esto se uso la la base de datos ***PathwayCommons*** para realizar la *Prueba Hipergeométrica* y el *Análisis de Sobre-Representación.*

Construcción de *Modelo Nulo.*

Con el fin de validar la estructura modular de la red, construimos un *modelo nulo* en el que se conservaron todos los nodos de la red MEF2C, así como el número de enlaces, pero los enlaces fueron reconectados aleatoriamente de acuerdo con el modelo de Erdös-Rényi [101] (sección 2.3.1 capitulo 2).

Redes transcripcionales de subtipos de cáncer de mama.

El cáncer de mama es la neoplasia maligna con mayor incidencia y mortalidad entre las mujeres del mundo [277]. Uno de los principales desafíos para su tratamiento es su naturaleza heterogénea (ver sección 1.3.2.1 capitulo 1), con manifestaciones que abarcan una gran cantidad de variantes clínicas, fisiológicas y de supervivencia, lo que da lugar a diferencias en las opciones terapéuticas disponibles [278].

Reconocer la importancia de comprender las relaciones genéticas podría ser determinante para establecer un mejor paisaje de la patología. En estos términos, analizar la tecnología de alto rendimiento para estudios genómicos puede ayudar en este empeño. Dado que la mayoría de los procesos desregulados no se han estudiado de manera integral, un enfoque holístico como el proporcionado por la teoría de redes resulta atractivo, ya que la estructura de red observada en redes biológicas tiene funciones funcionales ha sido demostrada en otro lugar [69, 210, 279, 280].

Cáncer de mama: una enfermedad heterogénea.

La heterogeneidad del cáncer de mama se puede rastrear hasta el nivel genético y molecular. La subtipificación molecular proporciona una herramienta útil para clasificar los tumores mediante la identificación de patrones comunes en su expresión genética. Se han desarrollado varios algoritmos que permiten la clasificación de muestras de cáncer mama en *subtipos moleculares* utilizando diferentes plataformas tecnológicas [281, 282]. Estos esquemas devuelven una clasificación general en términos de cuatro subtipos moleculares principales: *Luminal A*, *Luminal B*, *HER2+* y *basal.*

Luminal A. Alrededor de la mitad del total de los casos de cáncer de mama corresponden a tumores *luminales A* [283]. Estos tumores, a menudo son positivos para el receptor de estrógeno (*ER*) y negativos para el *receptor de tirosina quinasa* o *ERBB2* (HER2-), también presentan sobreexpresión de genes regulados receptor de estrógeno. Este subtipo generalmente tiene el mejor pronóstico [284] y las tasas de recurrencia más bajas [285, 286].

Luminal B. A pesar de que este subtipo tiene un patrón de expresión similar al de la luminal A, se caracteriza por una mayor variabilidad en la expresión de el *ER* y una mayor expresión de genes proliferativos, también se han encontrado en este subtipo, mutaciones asociadas con *TP53* e inestabilidad genética. Alrededor del 20 % del total de tumores de cáncer de mama se corresponden a este fenotipo [287], que tiende a tener un pronóstico más precario que el tumor *luminal A* [288].

HER2 enriquecido. Este subtipo intrínseco se caracteriza por la sobreexpresión del receptor *ERBB2* (HER2+), que está asociado con la amplificación del nivel cromosómico [289]. Estos tumores son negativos para los receptores de estrógeno y progesterona y tienen un pronóstico más precario que los de los subtipos luminales [290].

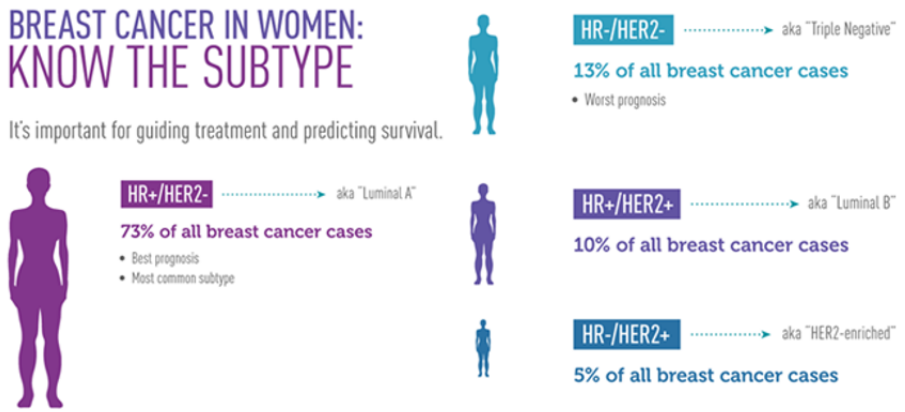


Figura 4.7: Esquema que representa los cuatro tipos de cáncer de mama su porcentaje en casos de .

Basal. 20 % de los tumores de mama son basales y la mayoría de *triple negativo* pertenecen a este subtipo. A diferencia de los subtipos descritos anteriormente, los tumores de tipo basal tienen una subexpresión de los receptores de estrógeno, progesterona y *ERBB2*. Estos tumores también están asociados con una mayor inestabilidad genética, son más agresivos y presentan el peor pronóstico. La mayoría de los tumores *BRCA*-1 relacionados con mutaciones pertenecen a este subtipo [285, 288, 291, 292, 293] y tienen un perfil de expresión génica similar al *epitelio mamario basal*. Es importante resaltar que el subtipo basal no se puede tratar con terapia hormonal convencional o anticuerpos monoclonales [281]. Por lo general, el tratamiento para pacientes con este subtipo incluye cirugía, radioterapia y quimioterapia. Esta es una razón importante para estudiar los patrones de regulación a nivel genómico de los genes asociados con la aparición de tumores de este subtipo. Asimismo, dado que los tumores basales no presentan *blancos* conocidos para terapia dirigida, la búsqueda de *módulos funcionales* que permitan focalizar y dirigir la busqueda experimental a un grupo reducido de moléculas que puedan

ser *blancos* resulta atractiva.

Las variaciones fenotípicas entre los subtipos de cáncer de mama surgen debido a las diferencias en sus programas de regulación transcripcional subyacentes [69]. Dichas diferencias en la estructura global de las redes de regulación específicas de cada subtipo, pueden reflejar diferencias en escalas de regulación más bajas, en particular en presencia de módulos funcionales subyacentes. Por lo tanto, analizar tales estructuras puede ser útil para comprender la organización particular de algunos procesos a lo largo de todo el sistema, y puede conducir a encontrar algunos indicios sobre la organización particular a nivel local [294].Así entonces, la estructura modular puede ser muy útil en el estudio de redes transcripcionales inferidas a partir de datos de cancer de mama, ya que un *módulo funcional* modela un conjunto de genes co-regulados que participan conjuntamente en un proceso biológico en este fenotipo [80, 81, 170, 195, 210, 215, 222, 223].

Así entonces para el caso de estudio de las redes transcripcionales asociadas a subtipos de cáncer de mama, se usaron redes inferidas partir de un conjunto de 493 muestras (perfiles de expresión de microarrays) de cáncer de mama [69]. Y usando la metodología expuesta en la sección 4.1, se detectó con *InfoMap* (sección 4.1.1) la estructura modular en la cada uno de los componentes conexos de dichas redes de regulación transcripcional. Una vez que se detectados los módulos (comunidades), se analizó si estos esran funcionales, es decir, si estban asociados estadisticamente a procesos biológicos específicos. En la figura 4.9 se muestra un flujo de trabajo completo para este estudio.

Inferencia de redes a partir de microarreglos de cáncer de mama.

Para este caso de estudio, utilizamos cuatro redes transcripcionales asociadas a cada uno de los subtipos moleculares de cáncer de mama, previamente descritas y analizadas por de Anda *et al.* [69]. Estas redes se infirieron a partir de un conjunto de 493 perfiles de expresión de microarreglos de muestras de cáncer de mama procesadas en la plataforma Affymetrix HGU133A. Se usó el algoritmo ARACNE [60] (sección 1.2.3) para calcular correlaciones entre pares de genes, las cuales se establecieron con el calculo de *Información mutua (MI)* a partir de los niveles de expresión medidos experimentalmente por estos microarreglos (figura 4.8).

Además, de forma complementaria, realizamos otros tres métodos para inferir las redes basados en medidas de correlación lineal: Pearson, Spearman y Kendall, sobre los mismos datos experimentales. Y asimismo, re-inferimos la redes mediante el calculo de *Información mutua (MI)* usando los algoritmos *CLR* y *MRNETBNetwork Science* mediante el paquete de R ***minet*** [295].

El conjunto de datos experimentales se recopiló de la base de datos *Gene Expression Omnibus* (GEO) a partir de los números de acceso GSE4922 [296], GSE1456 [297], GSE7390 [298], GSE1561 [299], GSE2603 [300] , GSE2990 [301], GSE9574 [302],

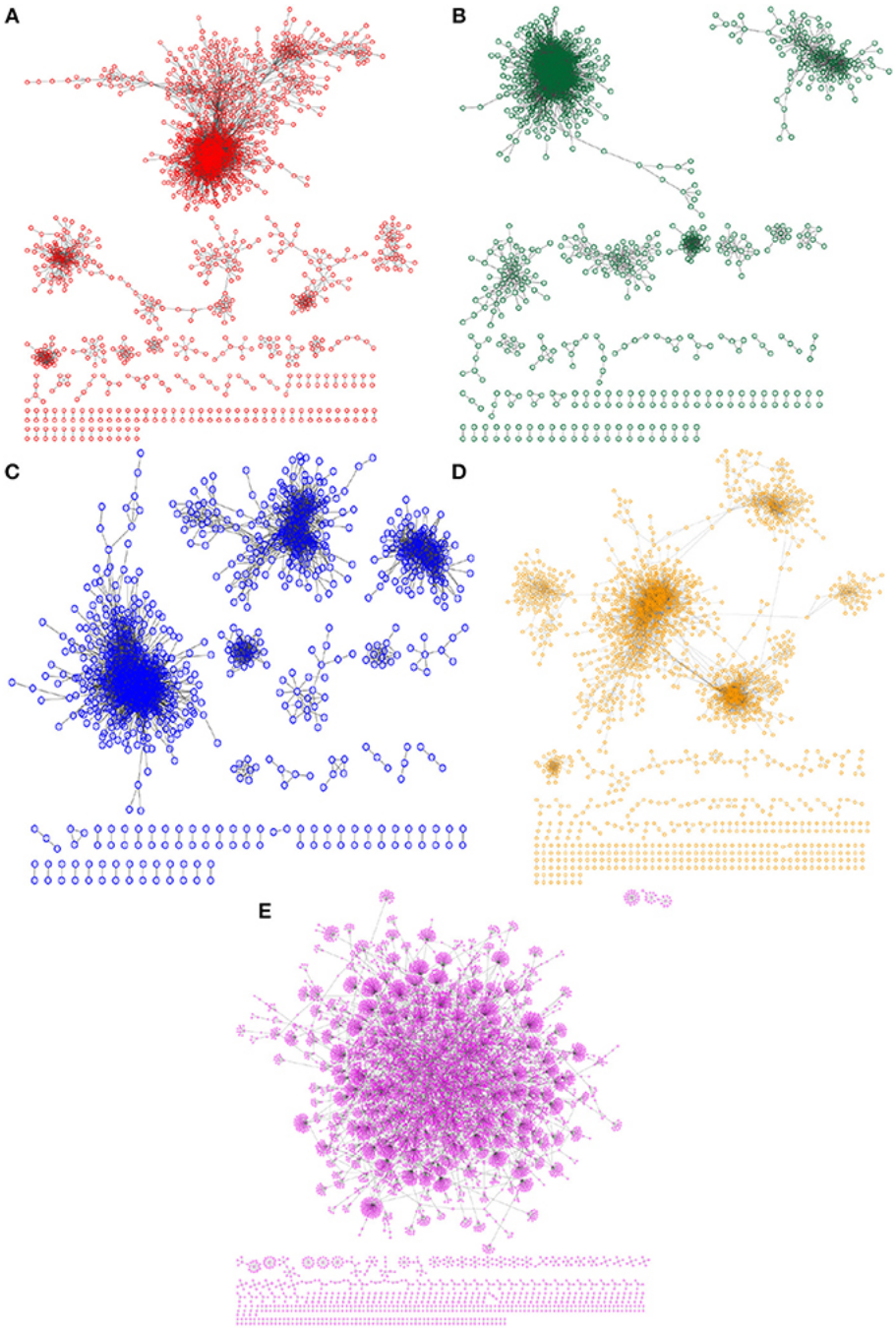


Figura 4.8: Redes asociadas inferidas a partir de biopsias de cáncer de mama. Se muestran las diferencias topológicas de las redes asociadas a los cuatro subtipos moleculares y la red de tejido sano.

GSE15852 [303], GSE6883 [304] y GSE3494 [305]. El preprocesamiento de datos se realizó siguiendo una linea de trabajo para *Robust Multi-array Average* [306], de la misma manera que Tovar *et al.* [307]. Las muestras de cáncer de mama se clasificaron usando el algoritmo ***PAM50*** [308]. La clasificación PAM50 de este conjunto de datos se logró previamente en [279].

Análisis de Enriquecimiento funcional.

Una vez detectados los módulos, para la red de cada cada subtipo, observamos la cantidad de estos y su tamaño respectivo. También se observamos si los módulos se conservan a través de subtipos moleculares para evaluar la existencia de un proceso de regulación común en los programas de transcripción de cáncer de mama. La hipótesis es que diferentes arquitecturas de red determinan estructuras modulares específicas y, por lo tanto, el programa regulatorio tendrá asociados diferentes procesos para cada subtipo molecular.

Así, realizamos un análisis de sobrerrepresentación para cada módulo, recurriendo a pruebas hipergeométricas corregidas via FDR con HTSanalyzeR [309], eligiendo una *significancia estadística* por debajo del valor $p_v \leq 0.001$. Buscamos en categorías *Gene Ontology* (GO) generales (es decir, ramas principales en el árbol de ontologías) representativas de la red de regulación del subtipo basal. Nos enfocamos en este subtipo ya que no presenta alternativas terapéuticas, sino citotóxicas o quirúrgicas. Al identificar los procesos enriquecidos para sus módulos, analizamos cuán específicos son. Por otro lado, también observamos si un proceso (o procesos) aparecía en varios módulos o subtipos.

Una vez que se realizaron los análisis funcionales, estudiamos aquellos genes con las medidas de centralidades más altas (*Betweenness centrality, clustering coefficient, shortest path length, grado y PageRank*) en cada módulo, para evaluar con mayor precisión el papel de dichos genes en el mismo y, lo que es más importante, en los procesos biológico sobrerrepresentados. Finalmente, teniendo en cuenta los genes diferencialmente expresados, investigamos si estos patrones de expresión podrían activar o inhibir los procesos enriquecidos estadísticamente significativos. Evaluamos el grado de desregulación de dichos procesos mediante cálculos de *z-score* [310].

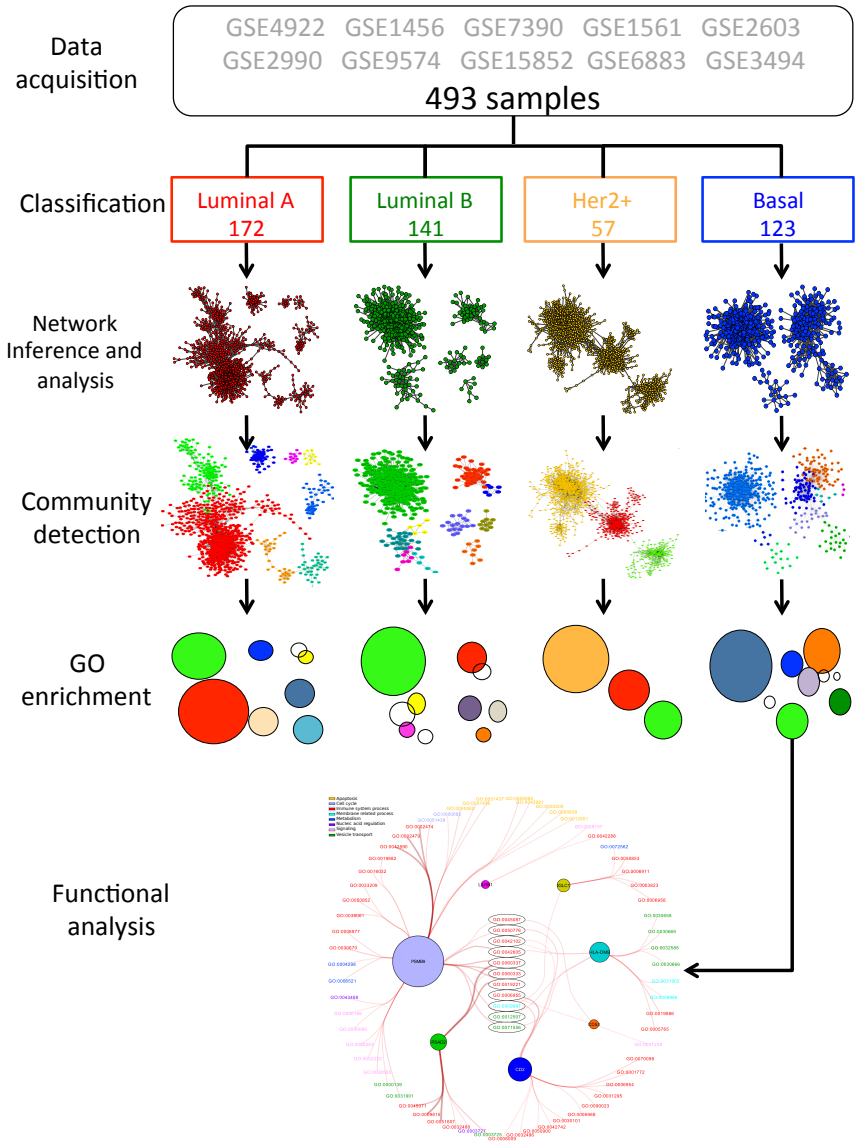


Figura 4.9: **Propuesta Metodológica para la búsqueda de módulos funcionales en redes de regulación genética asociados a los 4 subtipos moleculares de cáncer de mama**. El flujo de trabajo muestra la clasificación en subtipos moleculares de las muestras de biopsias de cáncer de mama; posteriormente la inferencia de redes y la partición en módulos de las mismas; para finalmente realizar un análisis de enriquecimiento el la base de datos Gene Ontology con el objetivo de asociar funciones biológicas a los módulos.

Submodularidad jerárquica en la red transcripcional del subtipo Her2 de cáncer de mama.

Como ya mencionamos, el subtipo HER2-positivo (*HER2+*) se caracteriza por la sobreexpresión del receptor *ERBB2*, es negativo para los receptores de estrógeno y progesterona y tienen un pronóstico más precario que los de los subtipos luminales [290]. Además es un ejemplo paradigmático del papel crucial que la *amplificación genética* y las *alteraciones en el número de copias* pueden tener en el cáncer de mama. En los casos de cáncer de mama HER2+, hay una alta tasa de amplificaciones en la *región Chr17q12-21*, también conocida como el *amplicón Her2*. Esta región incluye genes como TOP2, que codifica para Topoisomerase IIα, MED1, STARD3, GRB7, THRA, RARA, IGFPB4, CCR7, KRT20, KRT19 y GAST. La heterogeneidad de la expresión, las variaciones en el número de copias y las mutaciones involucradas en esta región tienen un fuerte impacto en el desarrollo del subtipo HER2+, resistencia a la terapia anti-HER2, progresión y pronóstico.

Como una consecuencia del caso de estudio anterior y dado que la estructura modular tiene, al mismo tiempo, una estructura jerárquica en su topología [74], decidimos analizar con mayor profundad la red del subtipo *Her2+*, dado que es la red más grande y su estructura modular podría contener información adicional a la obtenida con el análisis de modularidad funcional realizado. Así entonces, utilizamos la *mapaequation* jerárquica (sección 4.1.1.4), para encontrar submódulos anidados en la red del subtipo Her2+ (sección 4.2.2) y luego usando la metodología aquí propuesta (sección 4.1), identificamos si esos submódulos estaban asociados con una función biológica particular. Así, la red se analizó en tres niveles diferentes de modularidad: componentes conectados (islas), módulos en el componente más grande y submódulos en los módulos más grandes. La red se retomo del caso de estudio anteriormente expuesto (sección 4.2.2) y el análisis de enriquecimiento funcional fue más estricto (con un $p_v \leq 10^{-5}$), no sólo sobre los módulos sino también sobre los submodulos de la red.

Finalmente, es importante señalar que la apliación de la mtodología aqui presentada a los dierentes casos de estudio expuestos han sido publicados recientemente por nuestro grupo de trabajo [210, 215, 311].

Capítulo 5

Resultados y Conclusiones.

En este capítulo, presentaremos y discutiremos los resultados obtenidos de aplicar nuestra metodología para encontrar *módulos funcionales* en redes de regulación genética, aplicada a los casos de estudio descritos en el capítulo anterior. Asimismo expondremos la conclusiones del proyecto de doctorado.

Resultados de los casos de estudio.

En esta sección, presentaremos los resultados obtenidos tanto para la red transcripcional de MEF2C construida a partir del proyecto *Fantom4*. Así como los resultados obtenidos para las redes de cáncer de mama por subtipos, en particular la caracterización modular y jerárquica de la red del subtipo molecular *Her2*.

Red transcripcional de MEF2C.

En está sección, presentaremos los principales resultados del estudio de la red transcripcional MEF2C inferida por sitios de unión a ADN (TFBS) y sus principales características topológicas, la subestructura modular de la red, así como las funciones biológicas estadísticamente significativas encontradas en los módulos y finalmente se presentará la validación de estos hallazgos por contraste con un *modelo nulo*.

La red inferida del factor de transcripción MEF2C (Figura 5.1) está compuesta por 4543 nodos y 12422 enlaces (ver Tabla 5.1). Esta red muestra algunos genes altamente conectados (como GABPA, ELK4, CREB1, MYC o MEF2C), representados por círculos más grandes. Por ejemplo, MEF2C está profundamente implicado en el crecimiento y desarrollo muscular [312]. El análisis topológico de esta red se presenta en la Tabla 5.1. Es notable que MEF2C, HMGA2 y otros pertenecen a un grupo de TFs de gran relevancia conocidos como Reguladores Maestros, que corroboran resultados previos con respecto a los fenotipos de cáncer [273, 274].

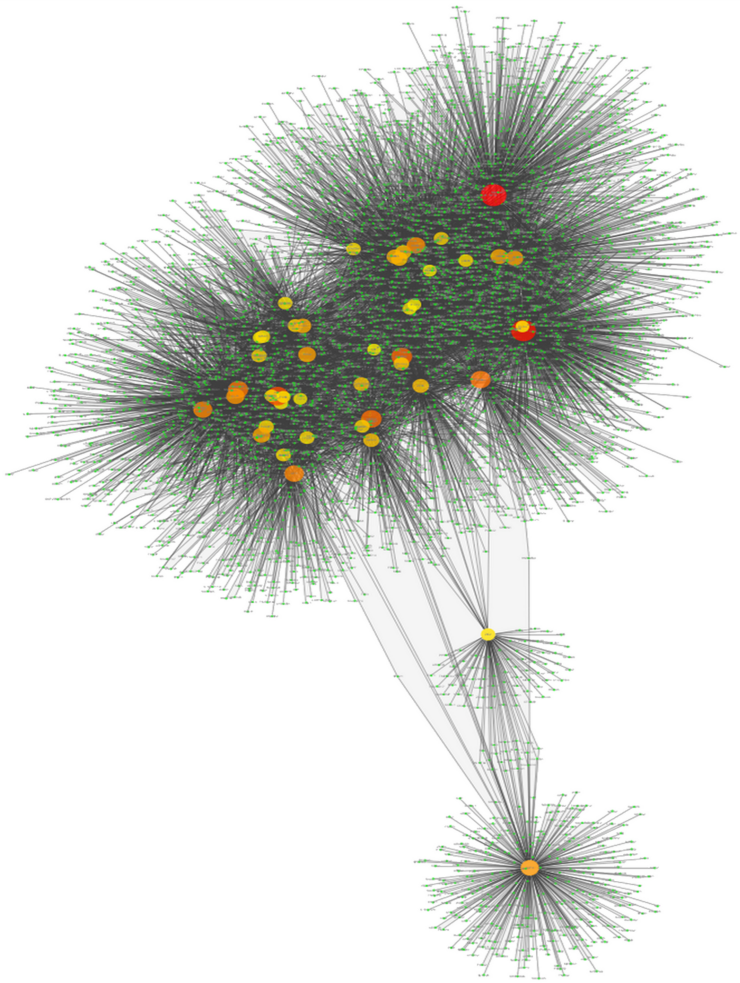


Figura 5.1: Red transcripcional de interacciones TFBS para el Factor de Transcripción MEF2C y sus blancos hasta el tercer nivel. En esta visualización, el color y el tamaño de los genes se representa según el grado de nodo (número de vecinos conectados a este gen particular): pequeños nodos verdes corresponden a genes apenas conectados; mientras que los nodos rojos y naranjas más grandes representan genes altamente conectados.

La estructura modular revela *sub-redes* que participan en procesos biológicos específicos.

La visualización modular de la red construida a partir de Infomap (Figura 5.2) contiene información sobre las comunidades que aparecen en la red. Estas comunidades

Número de modos n	4543
Número de aristas m	12422
Número de vecinos promedio $\langle k \rangle$	5.440
Coeficiente de Agrupamiento $\langle C \rangle$	0.250
Distancia mínima promedio entre nodos $\langle l \rangle$	3.170
Distribución de grado	$P(k) = 278.27 \times k^{-1.044}$
Distribución de *Clustering coefficient*	$C(k) = 1.907 \times k^{-1.042}$

Tabla 5.1: **Principales parámetros topológicos de la red transcripcional MEF2C.**

varían en la cantidad de moléculas y el flujo de información que las comunidades comparten entre ellos (para ver las listas de genes de cada comunidad, consulte el Material complementario 1). Por ejemplo, en esta red, las comunidades etiquetadas GABPA, ELF2 y NFYC son las más grandes y las que más comparten redes, en términos de información (Figura 5.2). Vale la pena mencionar que los círculos representan comunidades (en una perspectiva de grano grueso) que contienen varios nodos. El nombre de las comunidades viene dado por el nodo con el mayor PageRank [120] (sección 2.2.5.1.

Para aclarar la figura 5.2, expliquemos los aspectos gráficos de la misma tal como se realizó originalmente en [156]. El tamaño de cada módulo o "comunidad de genes" en el mapa refleja la fracción de tiempo que el caminante aleatorio[1] pasa dentro del mismo.

Como se puede observar, el tamaño de cada comunidad varía, sin embargo, el tiempo que pasa el caminante aleatorio dentro de una comunidad no es proporcional al tamaño (numero de nodos) de esa módulo sino al flujo de información dentro del mismo. Por ejemplo, la comunidad asociada con el gen GABPA incluye 643 genes y el caminante aleatorio mencionado anteriormente pasa el 14 % del tiempo de su caminata en esta comunidad: esta es la razón por la cual el nodo (módulo) GABPA es el más grande en la figura 5.2. Por otro lado, la comunidad NFYC incluye 309 genes, sin embargo, el caminante aleatorio pasa el 5,1 % del tiempo en esta comunidad. La pregunta que queda después de este análisis es si esas comunidades tienen un papel específico en los procesos biológicos, por ejemplo, si participan como un complejo multigenético durante eventos particulares en la célula.

[1] En general, con una "probabilidad de teletransportación" para "saltar" a otro nodo aleatorio en la red, (ver sección 4.1.1)

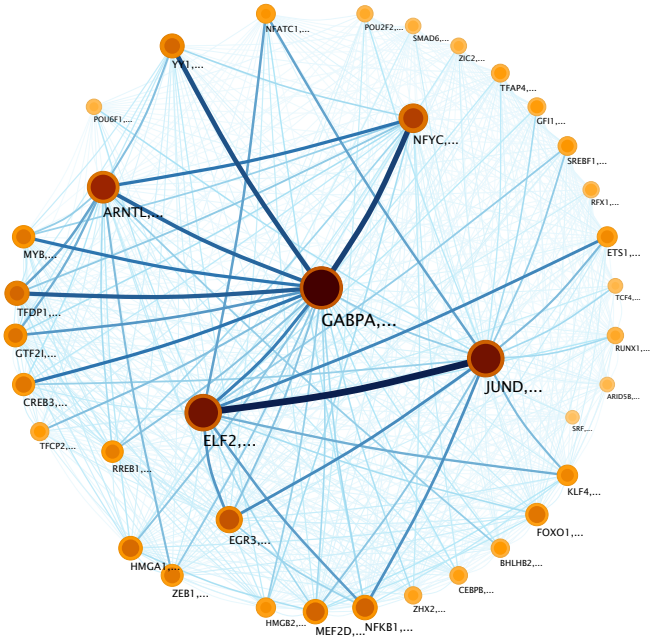


Figura 5.2: Estructura modular de la red transcripcional de MEF2C (figura 5.1). Los módulos se etiquetan con el nombre del nodo con el mayor PageRank (sección 2.2.5.1) dentro de la comunidad. Los nodos (que representan módulos) se representan de acuerdo con el tamaño de la comunidad y el flujo de información dentro de esa comunidad. En este sentido, los colores más oscuros corresponden a contenidos de información más grandes, mientras que los círculos más grandes representan comunidades más grandes. El grado relativo de flujo de información se representa en el ancho y el color de los enlaces entre módulos. El grosor de los bordes del módulo refleja la probabilidad de que un caminante aleatorio dentro del módulo siga una regulación (borde) a un gen fuera del módulo. Los enlaces ponderados entre comunidades representan *flujo de regulación*, con el color y el ancho de los bordes que indican el volumen de flujo. Por ejemplo, las líneas entre las comunidades JUND y ELF2 indican un flujo de información de regulación entre ellas. Estos enlaces revelan la relación entre las comunidades.

En las figuras 5.3 a 5.6, podemos ver mapas de calor de *p-values* (p_v) corregidos por FDR que muestran un enriquecimiento estadísticamente significativo de las vías biológicas dentro de comunidades específicas en la red. La intensidad del color es proporcional (*Z-score*) a $-\log_{10}(p_v)$. Las cifras 5.3 - 5.6 representan los procesos que se enriquecen

en comunidades relacionadas con la señalización celular (5.3), el ciclo celular (5.4), la expresión génica (5.5) y metabolismo (5.6).

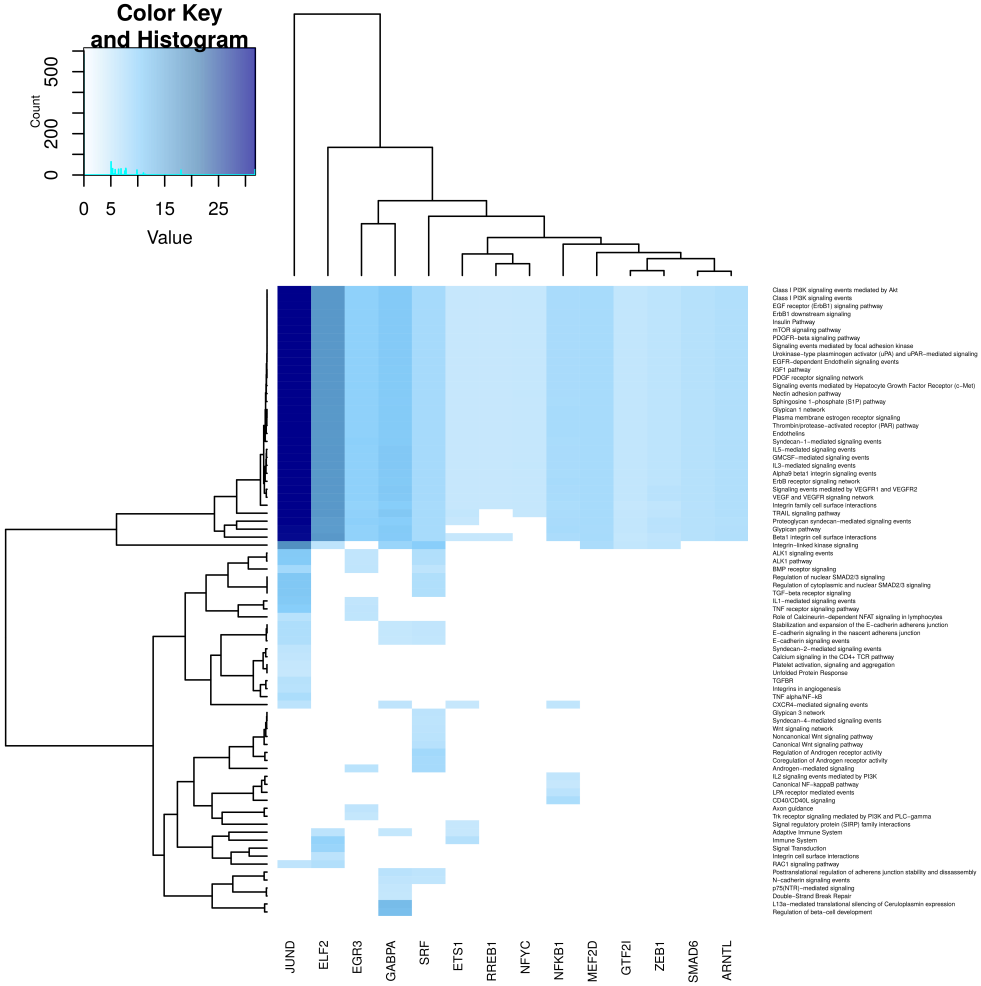


Figura 5.3: Mapa de calor que representa el enriquecimiento en las vías relacionadas con procesos de *señalización celular*. La intensidad del color es proporcional al $-log(p_v)$. Los puntos más oscuros corresponden a éxitos estadísticamente significativos.

En la figura 5.3, que se refiere a los procesos de señalización celular, se puede observar un fuerte enriquecimiento en las comunidades JUND y ELF2. También se destaca el hecho de que varios procesos se enriquecen en todas las comunidades (parte superior del mapa de calor). Los procesos más específicos se enriquecen en las comunidades GABPA, EGR3 y ETS1. Podemos observar que los procesos relacionados con el sistema inmune

también se enriquecen en la comunidad NFKB, lo que refuerza el hecho de que el gen NFKB participa en las respuestas del sistema inmune. Finalmente, vale la pena mencionar que muchos procesos relacionados con la señalización celular están ampliamente enriquecidos en el móudulo de SRF, a pesar de que es relativamente pequeño (17 genes).

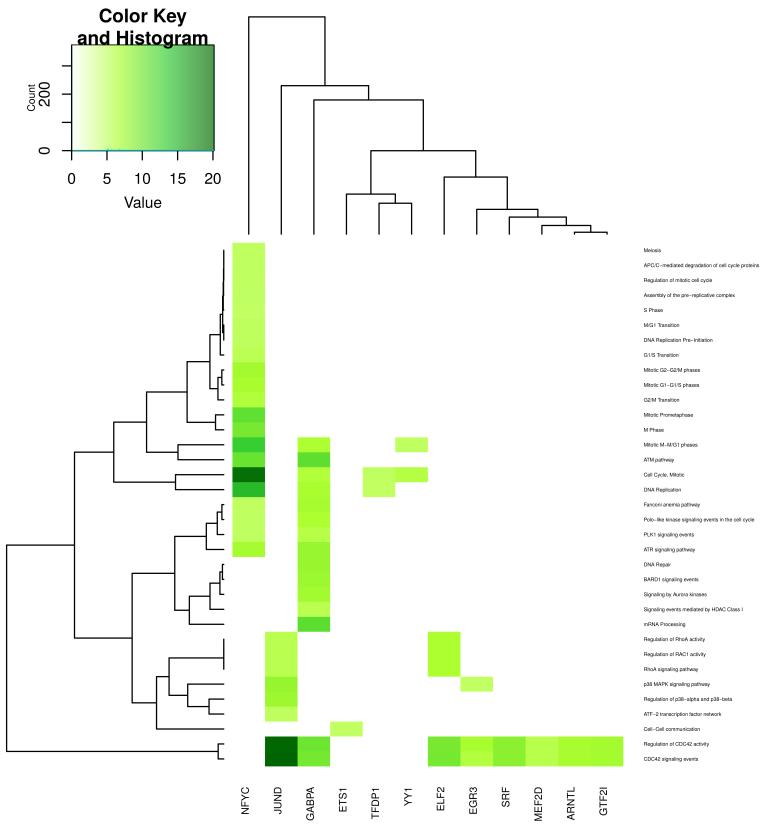


Figura 5.4: Mapa de calor que representa el enriquecimiento en las vías relacionadas con *ciclo celular*. Los módulos se etiquetan con el nombre de su molécula de con mayor PageRank (sección 2.2.5.1). El recuadro superior izquierdo muestra el histograma de *Z-score* y el color para el *p-value* de los procesos enriquecidos.

En el caso de la figura 5.4 se muestran procesos relacionados con el ciclo celular y la estructura del ADN, (a diferencia de la figura 5.3) menos de un cuarto de estos se enriquecen en la comunidad etiquetada como JUND; sin embargo, los procesos más enriquecidos de esta categoría pertenecen a este módulo como: **regulación de la actividad CDC42** ($p_v = 7.04 \times 10^{-21}$) y **eventos de señalización CDC42** ($p_v = 1.21 \times 10^{-20}$), respectivamente. Procesos como **ensamblado y degradación de las proteínas del**

ciclo celular solo se enriquecen en el módulo de NFYC, mientras que **reparación del ADN y procesamiento del ARNm** se enriquece exclusivamente en el módulo de GABPA. Nuevamente, un proceso muy específico como **cell-cell communication** solo se enriquece en el módulo ETS1.

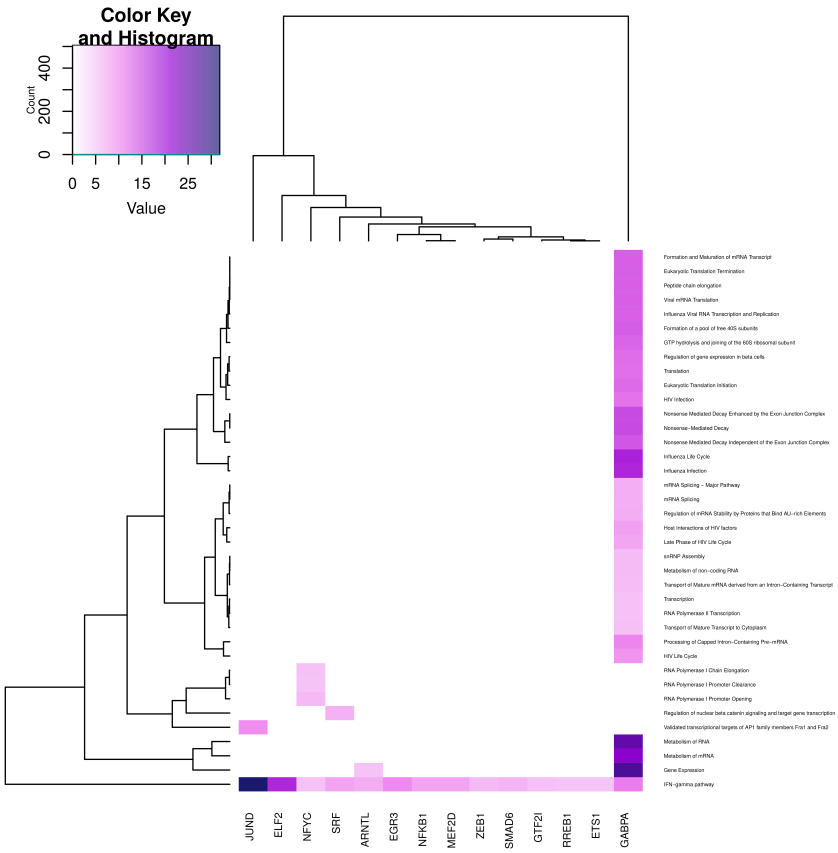


Figura 5.5: Enriquecimiento en las vías relacionadas con *expresión génica*. El recuadro superior izquierdo también describe un dendrograma que muestra distribuciones similares de *p-values* entre los procesos enriquecidos en las comunidades.

En cuanto a los procesos de expresión génica (Figura 5.5), todas las categorías, excepto las relacionadas con **transcription** están enriquecidas en la comunidad GABPA; en cambio, los eventos de transcripción se enriquecen en el módulo NFYC. Esta es otra instancia de la especificidad con respecto a la funcionalidad de las comunidades inferidas de topología. Por otro lado, **la via de IFN**γ se enriquece en todas las comunidades (parte inferior de la figura), lo que refleja la relevancia de este proceso. También vale la pena mencionar que el proceso de **regulación de nuclear** β**catenin signaling y**

target gene transcription nuevamente se enriquecen solo en la comunidad de SRF.

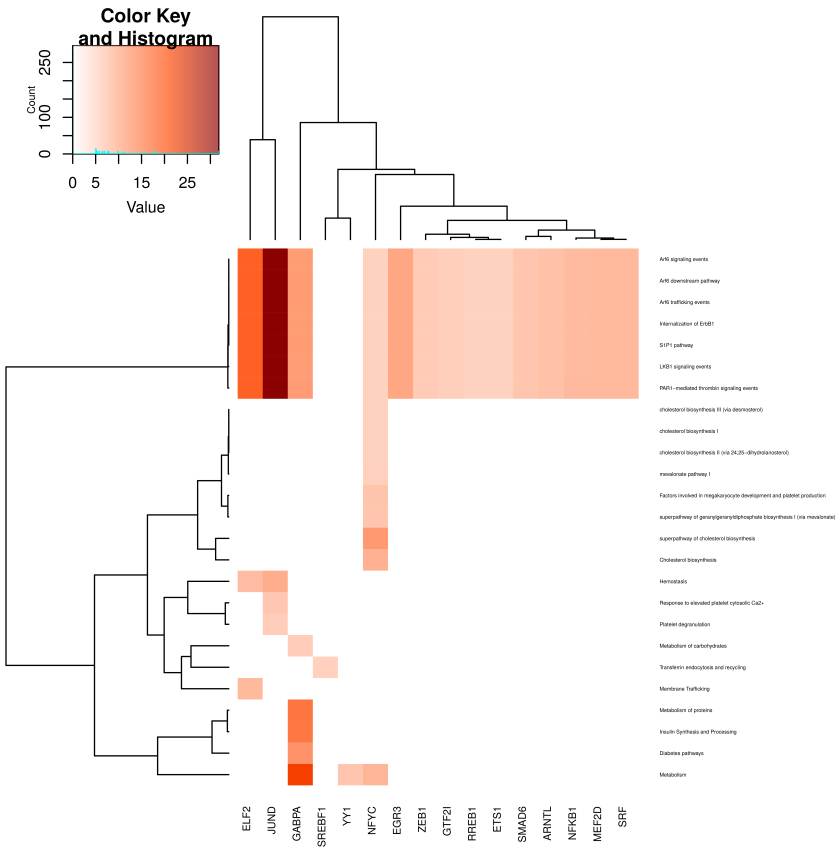


Figura 5.6: Enriquecimiento en las vías relacionadas con *metabolismo y procesos de transporte celular*.

Al observar Figura 5.6 (procesos metabólicos), un subconjunto de esos procesos se enriquece en casi todas las comunidades. Estos están más relacionados con **señalización Arf6** y procesos metabólicos de internalización. Estos procesos están altamente enriquecidos en la comunidad JUND. Además, los eventos relacionados con la diabetes y el metabolismo de proteínas se enriquecen en la comunidad GABPA. Los eventos más específicos, como **endocitosis y reciclaje de Transferina** solo se enriquecen en el módulo SREBF1; El módulo NFYC es específico para 8 procesos, principalmente relacionados con el metabolismo del colesterol y la producción de plaquetas y, finalmente, el tráfico de membranas se enriquece específicamente en la comunidad ELF2.

Podemos ver que los módulos están asociados con procesos biológicos específicos, como lo dan sus respectivas vías, es decir, tales vías se enriquecen en esa comunidad.

En algunos casos, los procesos están asociados de manera única con una comunidad específica, mientras que en otros, un proceso biológico dado puede enriquecerse en varios módulos.

La estructura modular es validada por un *modelo nulo*.

Para validar la estructura modular de a red transcripcional para interacciones TFBS para los factores de transcripción MEF2C se construyó *modelo nulo* (NMN) mediante una red Erdöos-Renyí con los mismos nodos y aristas, pero re-conectados de forma aleatoria. La figura 5.7 muestra dicha red (Figura 5.7 A) y la estructura modular de la misma (Figura 5.7 B) donde no se encontraron módulos. Por lo tanto, ninguna categoría de vías resultó significativamente enriquecida.

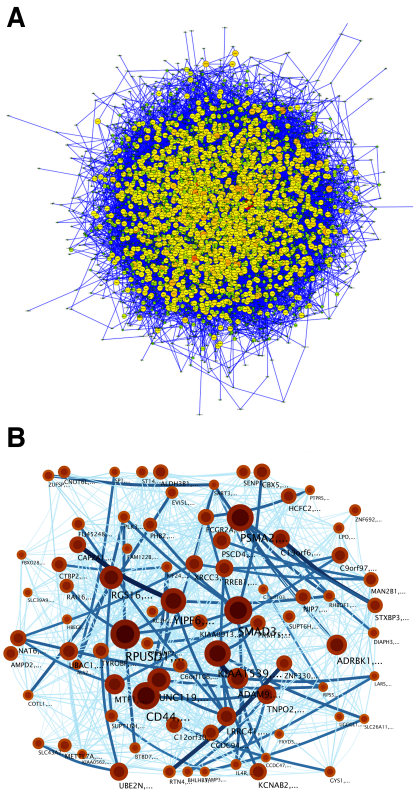


Figura 5.7: **Panel A.** Modelo nulo construido para la red de MEF2C. **Panel B.** Estructura modular del modelo nulo.

Finalmente, es importante señalar que los principales resultados de este caso de estudio han sido publicados por nuestro grupo de trabajo [210].

Redes transcripcionales de subtipos de cáncer de mama.

La estructura modular es específica para cada subtipo y está asociada a procesos biológicos específicos.

La tabla 5.2 resume brevemente algunas de estas diferencias estructurales de las redes de los los 4 subtipos de moleculares de cáncer de mama, inferidas previamente por de Anda *et al.* [69]. Dichas redes representan diferentes programas transcripcional asociados a los subtipos y cada una refleja diferentes arquitecturas de red transcripcional.

Subtipo	módulos enriquecidos	nodos en el componente más grande	conexiones en el componente más grande
Luminal A	8	930	8,535
Luminal B	5	555	8,476
Basal	7	523	7,181
HER2	3	1,649	9,108

Tabla 5.2: **Parámetros de la estructura modular de subtipos moleculares de cáncer de mama.**

De esta tabla, es posible observar diferencias entre las arquitecturas para cada subtipo de red. Teniendo en cuenta estas diferentes topologías, observamos que cada red transcripcional presenta un patrón de modularidad característico, es decir, la disposición de la regulación génica de los módulos y las reglas de conectividad para cada subtipo son únicas.

La tabla 5.2 también muestra que la cantidad de módulos para cada subtipo es diferente. En la figura 5.8, se muestra una visualización de cada red transcripcional, con los nodos coloreados por módulo. Una pregunta pendiente es si esos módulos tienen o no un papel funcional dependiente del fenotipo específico en el programa regulatorio de cada subtipo molecular de cáncer de mama.

Con el método de detección de módulos utilizado en este trabajo, todos los genes de la red se clasifican en un sólo módulo, cada uno. Por esta razón, tenemos módulos con procesos de Gene Ontology asociados, pero también módulos que no tienen categorías enriquecidas estadísticamente significativas.

En el caso de la inferencia de redes por *Información Mutua*, es necesario establecer

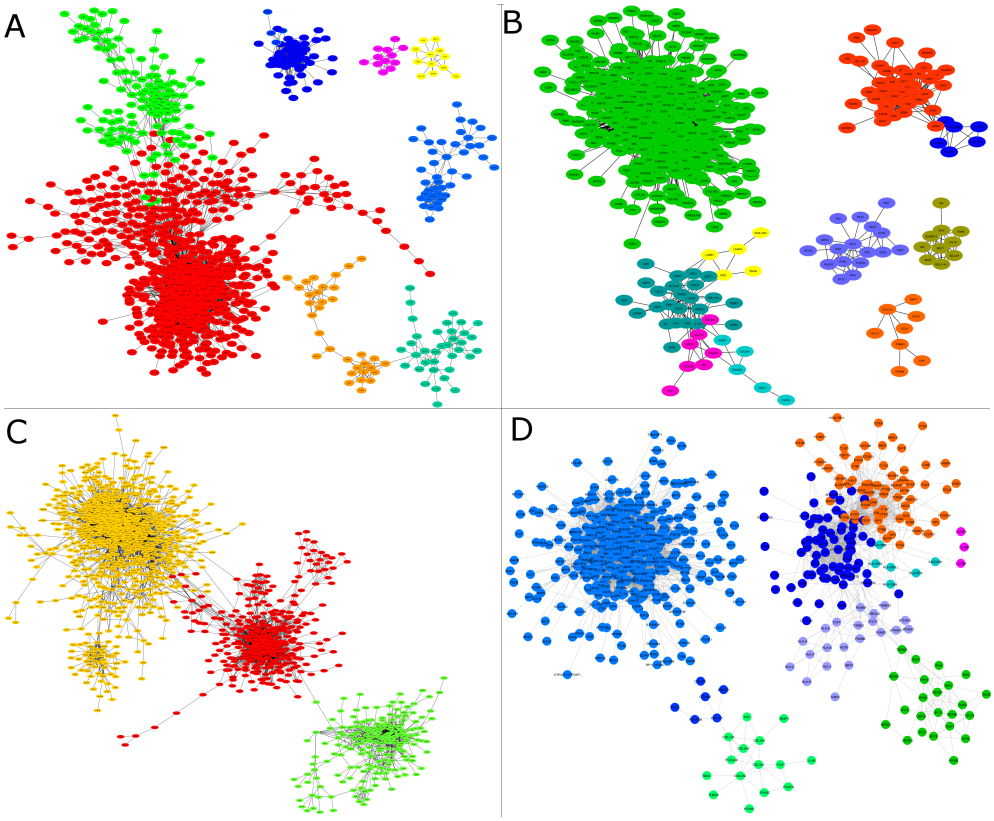


Figura 5.8: **Módulos en redes para cada subtipo molecular de cáncer de mama**. A) Luminal A; B) Luminal B; C) Her2 + y D) Subtipo basal. Los nodos que pertenecen a una comunidad tienen el mismo color.

un umbral para interacciones válidas. Para tener redes grandes pero *sparse*, se necesita un valor umbral estricto de información mutua. En [69] se demostró que las propiedades globales de las cuatro redes para los subtipos moleculares de cáncer de mama se conservan en una amplia gama de tamaños de red (3 órdenes de magnitud). Por este motivo, realizamos el algoritmo de detección de módulos y el enriquecimiento funcional con las mismas redes que los publicados en [69], es decir con $10,000$ enlaces correspondientes a los mayores valores de *Información Mutua* (pesos).

En la tabla 5.3, se puede observar cómo cada subtipo molecular tiene un conjunto único de módulos enriquecidos, con respecto al tamaño de los módulos y la cantidad de los mismos, así como también la cantidad de procesos enriquecidos para cada módulo. Vale la pena mencionar que el número de procesos enriquecidos no está directamente relacionado con el tamaño de los módulos (a partir de ahora, nombraremos los módulos de acuerdo con su nodo con mayor PageRank con la nomenclatura NombreGen_{comm}).

Por ejemplo, LUZP4$_{comm}$ para el subtipo luminal A está compuesto por 805 genes, sin embargo, solo 8 procesos están enriquecidos para este módulo. Por otro lado, ZFP36$_{comm}$ contiene solo 12 genes, pero 12 procesos están enriquecidos. Los procesos enriquecidos para ZFP36$_{comm}$ están relacionados con la respuesta al estímulo y la señalización.

Curiosamente, cada módulo en la red luminal A tiene un conjunto único de procesos enriquecidos. Esto podría estar relacionado con la estructura de la red (Figura 5.8 A), ya que casi todos los módulos están separados uno del otro. Este no es el caso para el resto de los subtipos moleculares. La lista completa de procesos enriquecidos por módulo para todas las redes de subtipos moleculares se puede encontrar en el material complementario 3.

El módulo asociado a COL5A2 está presente en cada subtipo molecular.

Anteriormente mencionamos que la estructura modular de cada red de subtipo molecular es diferente. Esto podría estar relacionado con el comportamiento específico antes mencionado observado en cada fenotipo. No obstante, el cáncer de mama tiene un núcleo común de características que pueden correlacionarse con el programa de regulación genética [280], comúnmente denominado *Hallmarks of cancer* [87].

En el caso particular de estas redes, queremos enfatizar el caso de las "comunidades COL5A2" (en negritas en la Tabla 5.3). En cada subtipo, identificamos un módulo en el que el gen con mayor PageRank es *COL5A2*, el gen de la proteína *colagenasa 5a2*, un componente integral de la matriz extracelular (MEC). Sin embargo, la composición del gen para COL5A2$_{comm}$ es diferente entre los subtipos moleculares. En la figura 5.9 A, se muestra un diagrama de Venn de la composición génica de COL5A2$_{comm}$ para cada subtipo molecular.

Se puede analizar una serie de características de este diagrama: cada módulo contiene una cantidad diferente de genes, pero más importante, solo hay tres genes que se comparten entre todas las redes de subtipos: *COL5A2*, *THBS2*, y *LUM*. *THBS2* codifica para la *trombospondina-2* un conocido comunicador célula-célula e inhibidor del crecimiento tumoral y la angiogénesis, que se ha asociado con cáncer gástrico y de mama [313, 314]. el gen *LUM* codifica para la proteína estromal *lumican*, que a su vez regula la organización de fibrillas de colágeno; las variaciones genómicas de este gen se han asociado con cáncer de mama [315].

En cada subtipo, COL5A2$_{comm}$ está asociado a procesos similares, como se muestra en el panel B de la figura 5.9. Hay cinco procesos enriquecidos comunes para COL5A2$_{comm}$ a través de los subtipos: *organización de fibrillas de colágeno*, *matriz*

	Módulo	genes	procesos enriquecidos
Luminal A	LUZP4	805	8
	NFIC	125	20
	COL5A2	**53**	**21**
	CD2	33	8
	TYROBP	36	3
	PLIN1	42	1
	KRT14	12	4
	ZFP36	12	12
Luminal B	LUZP4	464	15
	CD2	42	4
	COL5A2	**24**	**6**
	IFIT1	17	7
	IGKC	10	6
HER2+	CNR2	846	6
	LCK	370	83
	COL5A2	**196**	**29**
Basal	SLC4A4	390	7
	CD2	72	15
	CD53	65	3
	RSAD2	23	9
	PSMB9	21	38
	COL5A2	**15**	**16**
	IGLC1	12	6

Tabla 5.3: **Módulos enriquecidos encontrados en redes transcripcionales para subtipos moleculares de cáncer de mama.** Los módulos están etiquetados de acuerdo con sus genes con mayor PageRank (sección 2.2.5.1).

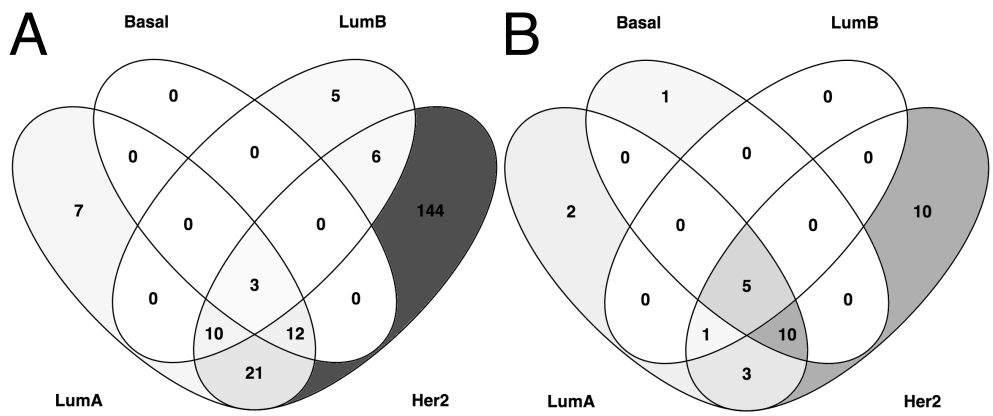


Figura 5.9: **Comunidades COL5A2: los procesos enriquecidos se comparten entre subtipos a pesar de que las composiciones genéticas son diferentes**. **A)** Diagrama de Venn que muestra el número de genes de COL5A2$_{comm}$ para cada subtipo molecular. Tenga en cuenta que solo se comparten 3 genes (*COL5A2*, *THBS2* y *LUM*). **B)** Procesos enriquecidos de COL5A2$_{comm}$ para cada subtipo molecular.

extracelular (ECM), *organización de matriz extracelular*, *constituyente estructural de matriz extracelular* y r*egión extracelular*. Esto puede implicar que los procesos relacionados con *ECM* son una característica común del cáncer de mama, a través del programa regulatorio de genes *COL5A2*, y esto es independiente del subtipo molecular.

Al observar la estructura modular en *COL5A2* a través los diferentes subtipos, pudimos discernir la existencia de un conjunto específico de procesos comunes a todos los subtipos de cáncer de mama: *desregulación ECM*. La aparición de procesos enriquecidos comunes es independiente de los genes presentes en cada red de subtipos, es una instancia clara de la solidez de los procesos cruciales adquiridos durante el desarrollo del cáncer de mama; la relevancia de la desregulación de la MEC puede tener injerencia en la malignidad del cáncer y se conoce bien su progresión [316, 317].

Además, dada la fuerte relación de la **MEC** con otros procesos durante el desarrollo del cáncer, como la *angiogénesis* (que se encuentra en el módulo COL5A2$_{comm}$ de HER2+), la inmunidad o la migración celular (revisada en [316]), este hallazgo adquiere más pertinencia. Los módulos COL5A2 tienen procesos relacionados con ECM enriquecidos por diferentes genes. Esos genes también contribuyen al enriquecimiento de otros procesos que pueden dar forma a paisajes específicos a través de subtipos moleculares.

Por ejemplo, en el subtipo HER2+ COL5A2$_{comm}$ tiene 10 procesos exclusivos enriquecidos significativamente, que incluyen *señalización TGF-β* y la *adhesión focal o*

angiogénesis. El enriquecimiento se produce a través de genes tales como TGF-β, *metaloproteinasas*, *colagenasas*, *VCAN* o *fibronectina* que aparecen en COL5A2$_{comm}$ de la red del subtipo HER2+.

TGF-β en particular, es bastante relevante para favorecer y promover un entorno *metastásico* [318]. TGF-β también está directamente activo en la *translocación de SMAD al núcleo*, promoviendo la expresión de varias moléculas relacionadas con ECM [319]. A pesar de las diferencias entre los genes para cada subtipo COL5A2$_{comm}$, los procesos relacionados con ***ECM*** aparecen consistentemente, y los procesos no comunes para cada subtipo pueden reflejar entornos particulares implicados en la biología del tumor para cada subtipo.

Dado que el patrón de expresión es crucial para inferir el efecto de una red de regulación genética, exploramos el perfil de expresión de los módulos COL5A2 en cada subtipo. Curiosamente, a pesar de que el mismo módulo contienen genes diferentes en cada subtipo, todos tienen un patrón de sobreexpresión, esto se observa claramente en la figura 5.10. En donde los genes están coloreados de acuerdo con sus niveles de expresión (rojo para sobreexpresado y azul para subexpresión).

Como se puede observar, el módulo COL5A2 en cada subtipo refleja que los procesos enriquecidos están exacerbados, lo que es consistente con el hecho de que los procesos relacionados con *ECM* están regulados positivamente, lo que corrobora las observaciones anteriores.

Análisis de los módulos del subtipo Basal y la disminución de la *apoptosis* mediante el módulo de *PSMB9*.

Con base en la asociación entre las estructuras modulares de estas redes y procesos biológicos, hemos podido observar que cada subtipo molecular tiene un paisaje funcional específico, que puede estar asociado con las características observadas en el entorno clínico. Con esto en mente, nos hemos enfocado en el estudio del subtipo molecular basal, el cual es el subtipo más maligno de cáncer de mama, con el peor pronóstico y alternativas terapéuticas más restrictivas. En la figura 5.11 A), podemos observar la estructura modular de dos componentes de la red del subtipo basal. Los colores representan los módulos y se puede notar que el segundo componente más grande y el módulo IGLC1$_{comm}$, que no está conectado al componente más grande, tienen procesos enriquecidos. El panel B) muestra los módulos enriquecidos (rellenos en color).

Un detalle de estos resultados se muestran en la figura 5.12, donde se observan las categorías de GO están involucradas para cada módulo. Los módulos se colorean de acuerdo con el código de color de la figura 5.11 y las etiquetas de los procesos enriquecidos se colorean dependiendo del tipo general del proceso.

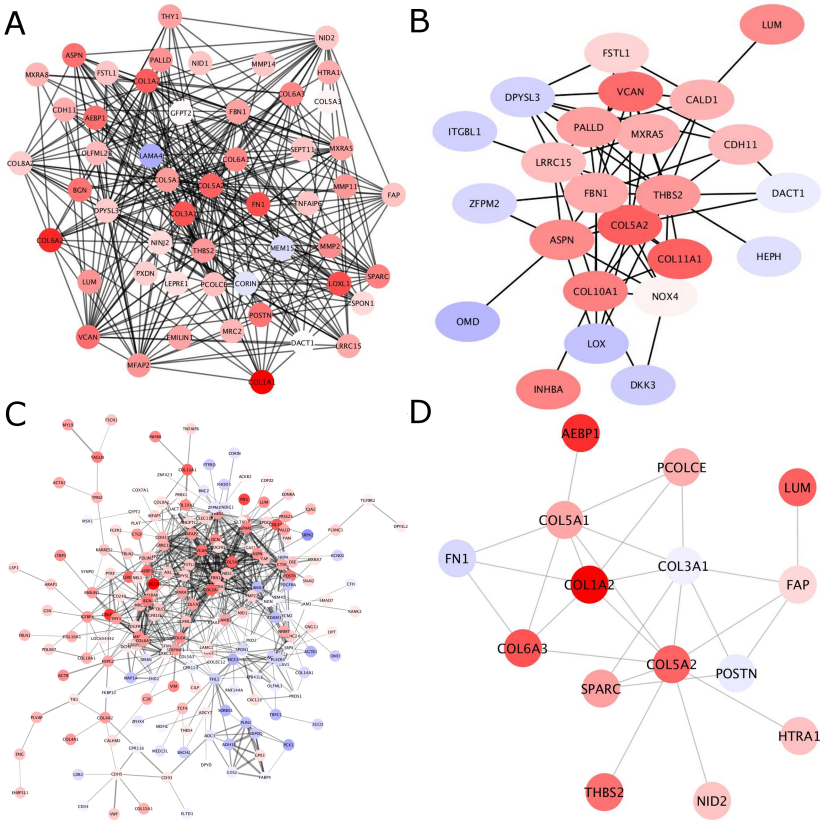


Figura 5.10: **Los genes en los módulos COL5A2 están mayormente sobreexpresados en todos los subtipos moleculares** Esta figura muestra la expresión de la firma de expresión de los genes que pertenecen a los módulos COL5A2 en **A)** Luminal A, **B)** Luminal B, textbfC) HER2+ y textbfD) subtipo basal de cáncer de mama.

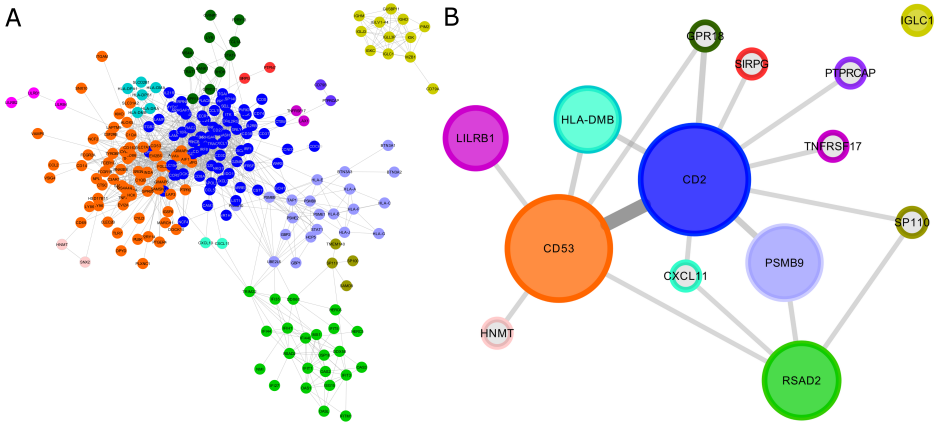


Figura 5.11: **Estructura modular en los componentes 2 y 6 del subtipo basal. A)** Los colores definen cada módulo. **B)** Flujo de información entre las comunidades. El ancho del enlace es proporcional a la cantidad de enlaces compartidos entre los módulos. Las comunidades a rellenas en color representan aquellas que están enriquecidas con alguna categoría de *Gene Ontology* (GO).

A partir de una inspección visual de la figura 5.12, es claro que la mayoría de las categorías enriquecidas de esos módulos están relacionadas con procesos del sistema inmune (etiquetas rojas). Sin embargo, algunos módulos tienen otros procesos enriquecidos, como es el caso de PSMB9_{comm}, que contiene solo 21 genes, pero se enriquecen con 38 categorías GO. Este es el único módulo en el que se pueden observar procesos enriquecidos relacionados con la ***apoptosis*** (etiquetas amarillas). En cuanto a las categorías enriquecidas de PSMB9_{comm} también observamos:

- 8 de los 38 procesos enriquecidos con PSMB están relacionados con *apoptosis* y no hay otro módulo que contenga procesos enriquecidos relacionados a este proceso biológico. Asimismo, los nombres amarillos en la Figura 5.12 solo están conectados a PSMB9_{comm}.
- PSMB9 también tiene otros 29 procesos únicos relacionados a este.
- 5 procesos de señalización también son únicos en PSMB9_{comm}.
- Las categorías relacionadas con *apoptosis* están involucradas en el complejo del *proteosoma* y su regulación.

También vale la pena mencionar que 40 de las 74 categorías GO enriquecidas pertenecen a procesos relacionados con el sistema inmune; 32 de ellos son categorías únicas

para módulos independientes. Esto puede implicar que, a pesar de que los módulos son conjuntos de nodos más conectados entre sí sobre el resto del componente, estas comunidades se comunican entre sí y presentan un programa regulatorio completo en el que los módulos específicos actúan individualmente como parte de un todo, al menos con respecto al *sistema inmune*. Los 11 procesos compartidos (8 de ellos relacionados con *sistema inmune*) en el centro de la figura refuerzan esta hipótesis.

La estructura modular de la red del subtipo basal, refleja no solo cómo los grupos de genes participan en un proceso funcional orquestado, sino también cómo los patrones de expresión de estos genes están de acuerdo con la dirección específica en que funciona dicho proceso, es decir, que se aumenta o disminuye. Como ejemplo de esto último, mencionamos el caso del módulo PSMB9_{comm}.

Como mencionamos anteriormente, el módulo PSMB9_{comm} está compuesto por 21 genes, pero está enriquecido en 38 procesos, en su mayoría relacionados con *apoptosis* y *sistema inmune*. Además, las centralidades de los nodos en este módulo revelan la relevancia de algunos genes en términos del flujo de información y una regulación coordinada de subprocesos. Por ejemplo, PSMB9, TAP1 y UBE2L6 son los genes con las más altas centralidades (*Betweenness centrality*, *clustering coefficient*, *grado* y *PageRank*). Debido a estas propiedades, su eliminación estos nodos divide la subred en dos partes: por un lado, el módulo de genes HLA fuertemente conectados, relacionados principalmente con el *Complejo Mayor de Histo-Compatibilidad* (MHC), y por el otro lado, el *complejo del Proteasoma*, relacionado estos nodos con apoptosis y ubiquitinación. Con base en estas mediciones de centralidad (entre otras), podemos argumentar que estos elementos están regulando de forma coordinada los procesos de inmunidad y la muerte celular.

También es importante el hecho de que todos los genes en el módulo PSMB9_{comm} están sobreexpresados. Teniendo en cuenta que la *muerte celular* y la *inmunidad* actúan coordinadamente en el subtipo basal, investigamos la predicción de activación o inhibición que estos procesos presentan, es decir, basados en el perfil de expresión del conjunto de genes, cuál es la dirección de cambio para una función determinada. En este caso, *muerte celular* e *inmunidad*. Para este propósito, utilizamos el *Análisis de Enfermedades y funciones* proporcionado por el *Ingenuity Pathway Analysis* de QIAGEN (IPA®, *QIAGEN Redwood City*, www.qiagen.com/ingenuity), que evalúa (por medio de un *z-score*) la coincidencia entre los patrones de regulación *up/down* observados y predichos, como se describe en [310, 317].

Coincidentemente, la función más inhibida predicha por el análisis es *muerte celular* (con *z-score* de -8), mientras que la función más activada, predicha por el análisis fue *infección viral* (con *z-score* de 7). Esto significa que el perfil de expresión de las moléculas relacionadas con ambos procesos tienen una expresión que se observa cuando los eventos relacionados con la muerte celular disminuyen y, de forma concomitante, los

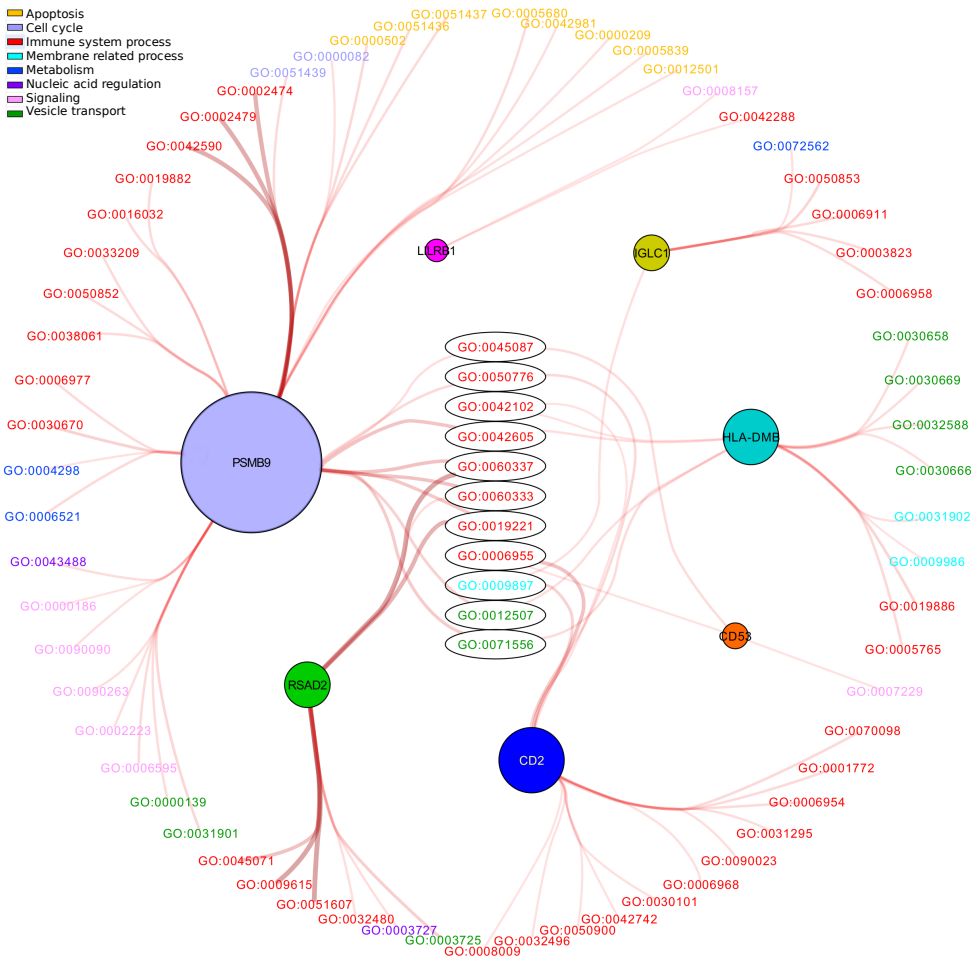


Figura 5.12: **Procesos de *Gene Ontology* (GO) asociados con módulos en la red de transcripción del subtipo basal de cáncer de mama.** En esta figura, los módulos se colorean de acuerdo con el código de color de la Figura 5.11. Estas comunidades están conectadas a las categorías de GO que están coloreadas de acuerdo con un proceso general.

genes que responden a una infección viral activan un esquema de defensa. La figura 5.13 muestra un subconjunto de moléculas que participan en ambos procesos y el efecto de predicción sobre ellas. Como se puede observar, la misma firma de expresión inhibe la muerte celular y, al mismo tiempo, parece activar genes de infección viral.

Finalmente, es importante señalar que los principales resultados de este caso de

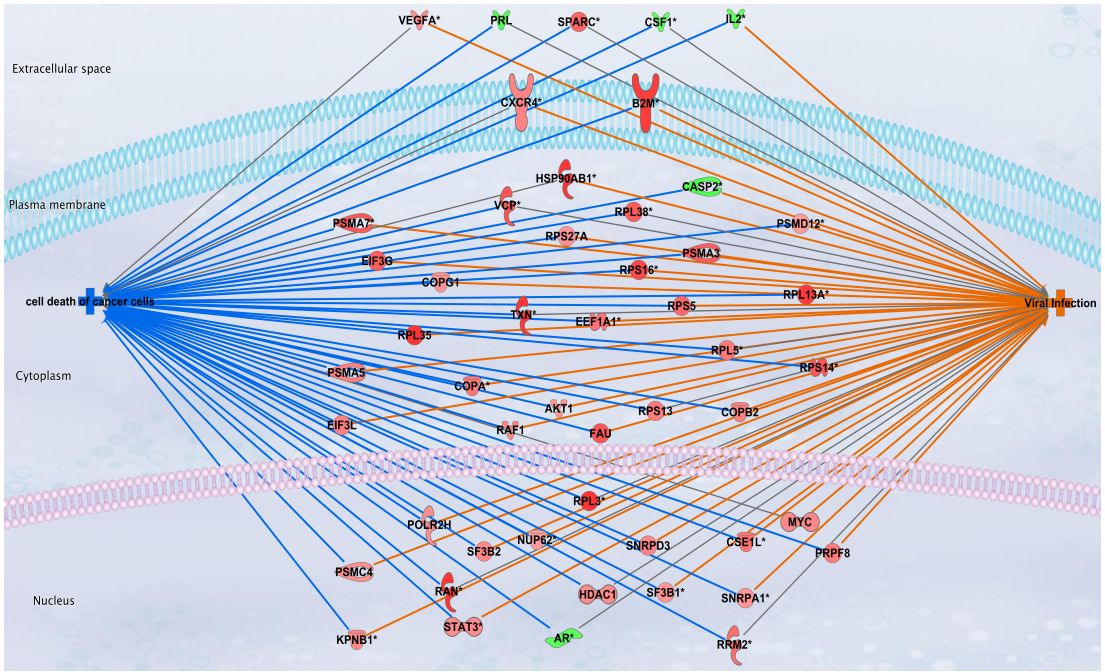


Figura 5.13: **Los procesos de muerte celular e infección viral están regulados de manera opuesta por la misma firma molecular del subtipo basal de cáncer de mama.** En esta figura, los genes se representan según sus niveles de expresión: rojo para sobreexpresado y azul para genes subexpresados. Las líneas entre las moléculas y procesos indican la función prevista de la molécula de acuerdo con su valor de expresión, la línea azul conduce a una inhibición del proceso; a su vez, las líneas naranjas representan la activación prevista. Los procesos de color de **muerte celular** y **infección viral** representan el mismo efecto predicho que las líneas.

estudio han sido publicados recientemente por nuestro grupo de trabajo [215].

Submodularidad jerárquica en la red transcripcional del subtipo Her2 de cáncer de mama.

A partir del análisis expuesto en la sección anterior, observamos que la red Her2 + está compuesta por unos pocos componentes grandes, que a su vez están divididos en más submódulos. Detectar dichos submódulos nos permite observar con más detalle los conjuntos de genes que pueden participar en algún proceso biológico.

La figura 5.14 **C)** y **D)**; muestran la estructura modular y submodular de la red Her2+. En la figura 5.14 panel B), es posible observar que los genes se agrupan en subconjuntos donde el patrón de expresión es similar. El grupo situado a la derecha está compuesto principalmente por genes subexpresados (coloreados en azul). Por otro lado, los genes sobreexpresados (coloreados en rojo) se agrupan en dos subconjuntos principales, donde los grupos dentro de esos subconjuntos también están compuestos por genes con patrones de expresión similares. A partir de una inspección visual, pequeños grupos en la parte central izquierda de la Figura 5.14 panel B) muestran sobreexpresión. Estos pequeños subconjuntos corresponden a módulos relacionados con la *regulación transcripcional* y *respuesta viral*. Ambos módulos serán ampliamente discutidos a continuación.

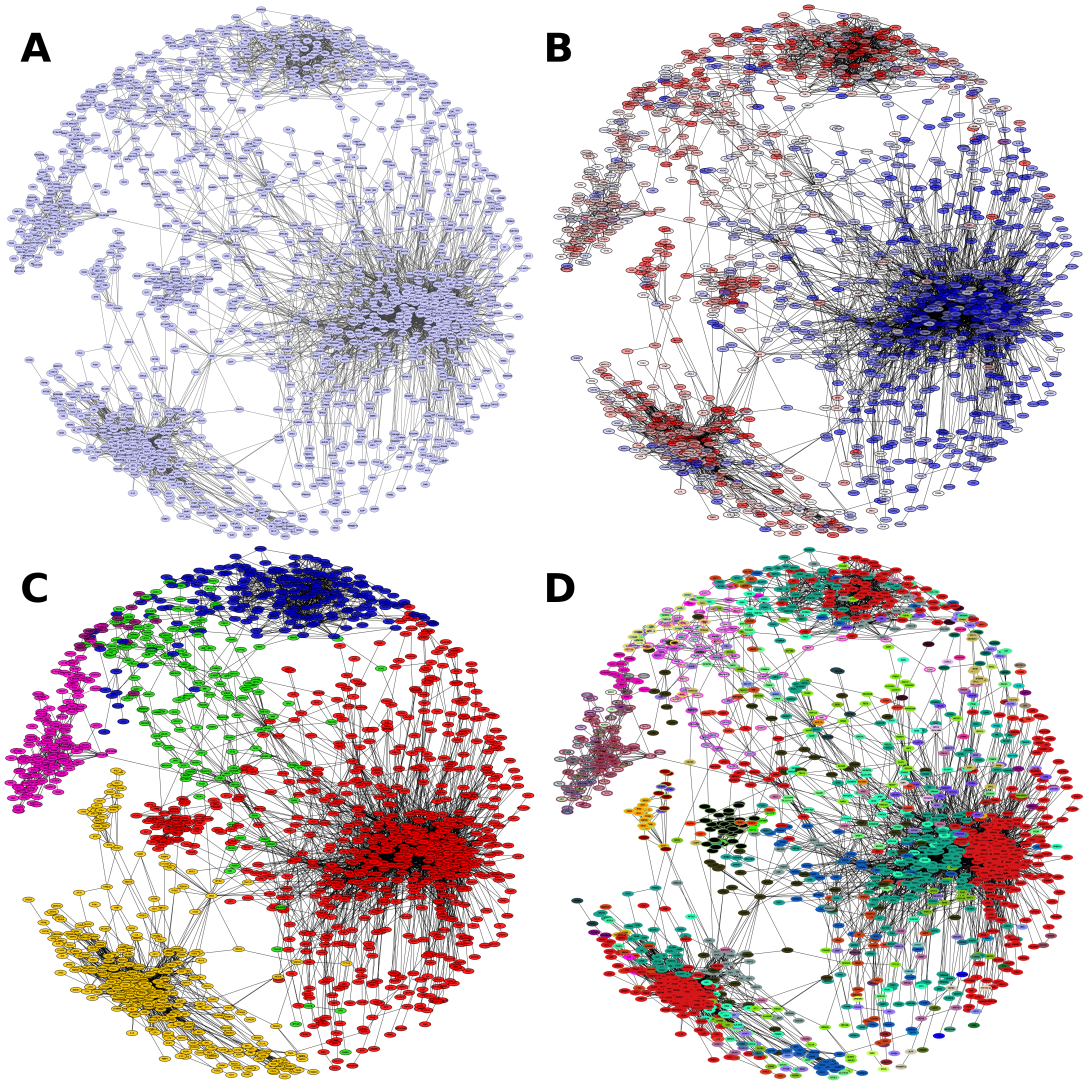


Figura 5.14: **Componente gigante de la red del subtipp Her2**. **A)** Se muestra la arquitectura de red de Her2. **B)** Genes expresados diferencialmente, los nodos rojos representan genes sobreexpresados, mientras que los nodos azules son genes subexpresados. **C)** Estructura modular de la red de cáncer de mama Her2+ **D)** Estructura submodular. Nótese que en B), los genes con un patrón de expresión similar se agrupan.

Enriquecimiento y análisis de expresión diferencial, revela la función biológica en submódulos.

Descubrimos que la estructura modular tiene una estructura particular. En primer lugar, los genes de la red tienden a agruparse de acuerdo con su nivel de expresión: los genes sobreexpresados están relacionados con genes sobreexpresados y asimismo los subexpresados.

Además, los módulos obtenidos también tienden a estar sobreexpresados o subexpresados, lo que sugiere una acción concertada de ellos. Además, el análisis de enriquecimiento de esos módulos muestra procesos importantes relacionados con conjuntos de genes específicos.

Entre los resultados más relevantes se encuentran: el módulo que contiene el gen de colágeno 5a2 (COL5A2), el cual está fuertemente relacionado con la matriz extracelular (ECM); el módulo CNR2 está relacionado con la regulación de la membrana plasmática y de micro ARN. Finalmente, un módulo sobreexpresado en su mayoría en el cual el gen con mayor PageRank es el protooncogén LCK, el cual está altamente relacionado con respuesta a infección viral, principalmente mediada por el módulo de genes OAS2. Este último resultado puede sugerir que el fenotipo HER2+ podría estar involucrado en una respuesta fisiológica viral.

Estos resultados destacan la importancia de estudiar el cáncer como una enfermedad basada en procesos, más que una basada en genes. Los hallazgos obtenidos aquí posiblemente no podrían haber sido observados sin la estructura de la red. Comprender la arquitectura de los fenotipos complejos transcripcionales puede ayudar a diseccionar los mecanismos detrás de la aparición de un fenotipo patológico.

Como se ha mencionado anteriormente, los módulos y submódulos se componen de genes con un patrón de expresión similar. Dado que la red presentada aquí se construye calculando la *Información Mutua* entre cualquier par de genes a lo largo de una gran cantidad de muestras, se espera que se observen las correlaciones más altas entre los genes que tienen un patrón de expresión similar en todas las muestras. Teniendo en cuenta esto último, tener módulos que contengan genes que participen en procesos similares resulta atractivo. Esa es la razón por la cual decidimos analizar los submódulos, buscando funciones más específicas para finalmente comprender el paisaje patológico en el fenotipo Her2+.

Además de lo anterior, tres de los módulos principales están estadísticamente enriquecidos en varios procesos biológicos de acuerdo con la base de datos *Gene Ontology* [184], estos no se superponen entre sí, lo que muestra que cada módulo está asociado a funciones biológicas muy específicas (ver figura 5.15). Los módulos y submódulos se etiquetaron de acuerdo con el gen con la centralidad de PageRank más alta [120] (sec-

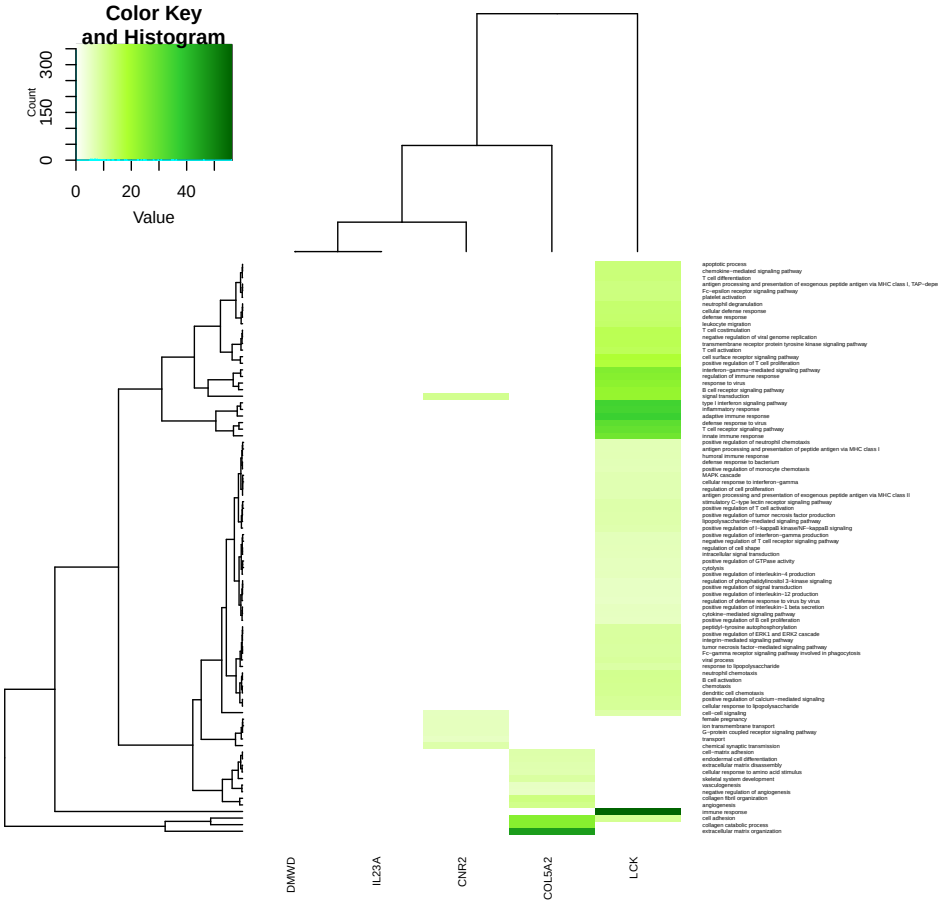


Figura 5.15: **Procesos biológicos enriquecidos por módulo en la red del subtipo Her2+ de cáncer de mama.** La figura muestra los procesos enriquecidos (renglones a la derecha) por módulo (columnas con nombres en la parte inferior). Nótese que el módulo LCK tiene la mayoría de los procesos enriquecidos, pero los procesos no enriquecidos en LCK se enriquecen en los demás módulos. El gradiente de color es proporcional al $-log(p_v)$ de los procesos enriquecidos.

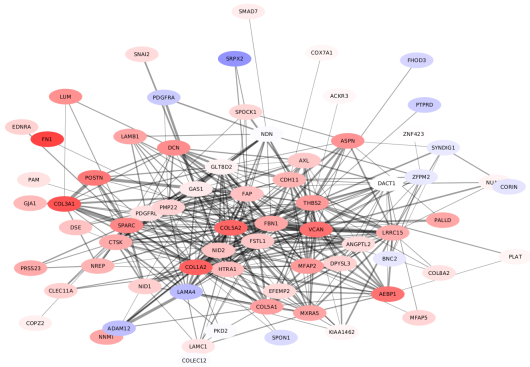


Figura 5.16: **Perfil de expresión del módulo COL5A2**. Este submódulo está compuesto principalmente por genes sobreexpresados. Interesantemente, los genes subexpresados (representados en azul claro) tienen un pequeño número de conexiones, en comparación con el número de enlaces que tienen la mayoría de los genes sobreexpresados. Nótese que los genes incluidos en este módulo son la Familia de Colágeno, además de FN1, FBN, VCAN, LUM y THBS2.

ción 2.2.5.1), es decir, cada módulo se denomina por su gen con PageRank más alto.

En lo que sigue, presentaremos los resultados más relevantes en los submódulos, basados en el análisis de enriquecimiento y patrones de expresión diferencial de dichos submódulos.

En la figura 5.14 panel C), el módulo azul en la parte superior de la red, el gen con el PageRank más alto es COL5A2. Este gen codifica para la proteína *Colágeno 5A2*, un participante clave en la conformación de la matriz extracelular. Este gen está muy sobreexpresado, lo que a su vez se ha observado en el cáncer de mama []. Ahora bien, el módulo COL5A2 está compuesto a su vez por submódulos. En el submódulo más grande, el gen con el PageRank más alto es nuevamente el gen COL5A2. Este submódulo está enriquecido para procesos relacionados con la *matriz extracelular*, la *adhesión celular* y la *organización de fibrillas de colágeno*. Esto es un claro indicativo de que la remodelación de la matriz extracelular es un participante clave en la formación del fenotipo en el subtipo de cáncer de mama Her2+, y asimismo es fundamental para *invasividad*, *migración*, *transición epitelio-mesenquima* (EMT) y otros procesos ampliamente observados en el cáncer. Vale la pena mencionar que los genes de este submódulo están sobreexpresados (Figura 5.16). Genes como LUM, la familia de colágeno, fibronectina, VCAN y otros relacionados con los procesos mencionados anteriormente también aparecen en este módulo.

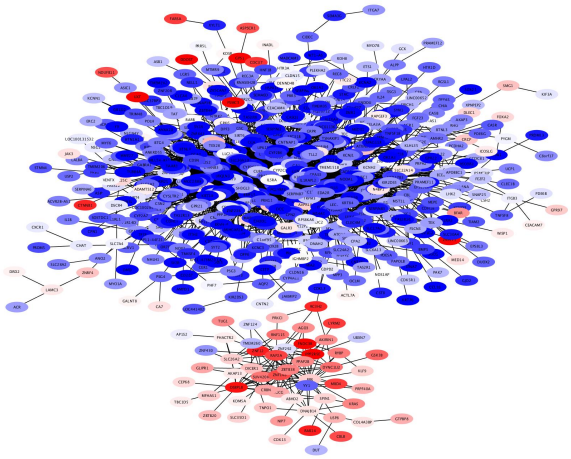


Figura 5.17: **Perfil de expresión del módulo CNR2**. En esta figura, es evidente el módulo sobreexpresado en la parte inferior; este módulo está enriquecido para procesos asociados a la membrana plasmática.

Submódulos CNR2 y la membrana plasmática.

En el caso del módulo CNR2, que es la comunidad más grande, compuesta por 763 genes, la mayoría de sus genes están sobreexpresados (módulo derecho en la figura 5.14, y figura 5.16) Interesantemente, el módulo CNR2 tiene 51 submódulos, dos de estos, el módulo ZBTB38 y los submódulos GPATCH4, contienen a su vez 5 y 4 submódulos, respectivamente. A partir del análisis de enriquecimiento de los submódulos, resulta notable que solo 3 módulos resulten enriquecidos: CNR2, PGLYRP4 y ZBTB38. A partir de estos enriquecimientos, las categorías relacionadas con la membrana plasmática se enriquecen.

Además, el submódulo ZBTB38 tiene otras características interesantes: es el único módulo de la comunidad CNR2 que está sobreexpresado en su mayoría. Además, una inspección más profunda de este subconjunto de genes revela que algunas de las moléculas más importantes para procesar la actividad de micro ARN aparecen allí: genes DICER y AGO3, que son elementos clave en el complejo RISC, que es el complejo que regula el gen expresión a través de micro-ARN. La figura 5.17 muestra este módulo, resaltando estas dos moléculas. Se ha demostrado que la regulación de micro ARN es crucial para el mantenimiento de la transición epitelio-mesenquima, pero también para los procesos mesenquima-epitelio. Por lo tanto, observar un módulo relacionado con el complejo de proteína que es responsable de la regulación gen-microARN, resulta relevante para una posible nueva opción terapéutica.

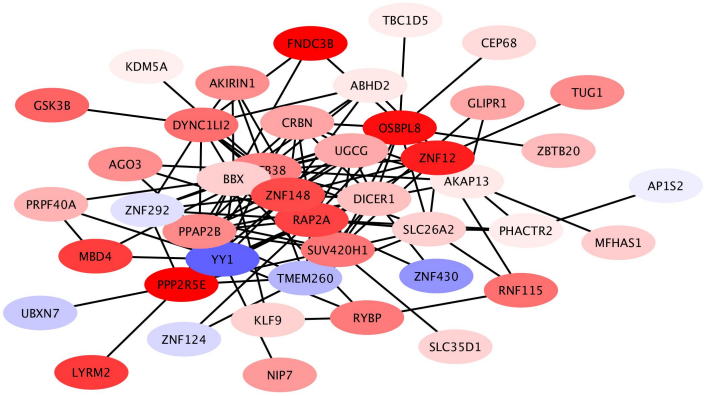


Figura 5.18: **Perfil de expresión del submódulo ZBTB38 del módulo CNR2**. Este módulo está sobreexpresado en la mayoría de sus genes. Nótese que los genes DICER y AGO3 (resaltados), que codifican las proteínas DICER y Argonnauta3, parecen estar sobreexpresados, lo que sugiere un papel importante de esta comunidad en la regulación gen-microARN.

Submódulos de LCK, respuesta viral y respuesta inmune celular.

En el caso del módulo LCK, compuesto por 371 genes, a partir de una inspección visual de la figura 5.19 panel **A**, es posible observar que la mayoría de los genes en el módulo están sobreexpresados. Este módulo se subdivide en 27 módulos más pequeños y uno de ellos, el módulo OSA2 (figura 5.19 panel **B**), a su vez se subdivide en tres submódulos secundarios. El submódulo LCK, el más grande del módulo LCK, está enriquecido para varios procesos relacionados con el sistema inmune. De hecho, casi todos los procesos enriquecidos en módulos LCK y submódulos se enriquecen para las categorías relacionadas con el sistema inmune. En el caso de la subcomunidad OAS2, este submódulo está altamente enriquecido para eventos relacionados con virus (figura 5.20).

Este último resultado no es la primera vez que se reporta. Como vimos en la sección anterior, varias moléculas que participan en la disminución de los procesos relacionados con la muerte celular, también promueven una respuesta viral. En este caso, se realizó un análisis de enriquecimiento funcional que tiene en cuenta el valor de expresión del gen, pero también el valor esperado que el gen debería tener en el contexto de un proceso particular. El análisis resultó en que los procesos con los *z-scores* más altos se relacionaron con la respuesta viral.

En la figura 5.20, se muestra una matriz en la que los tamaños de los cuadrados

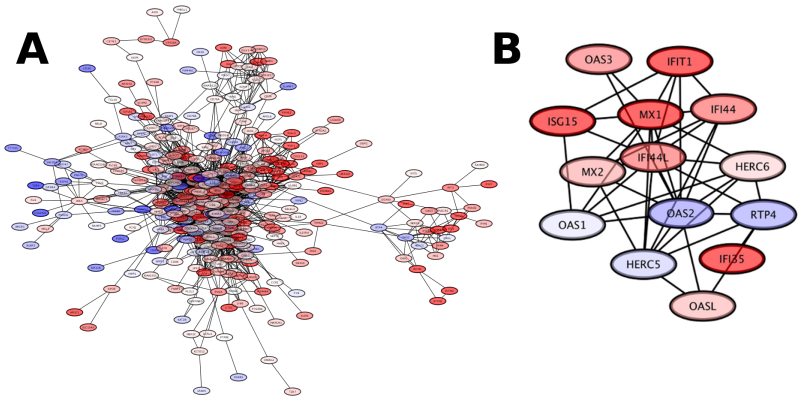


Figura 5.19: **Perfil de expresión del módulo LCK y respuesta viral**. A) El perfil de expresión del módulo LCK se muestra aquí. Observe la clara separación entre los pocos genes subexpresados a la izquierda de la figura. B) El submódulo en el cual el gen OAS2 es el de mayor PageRank, también está enriquecido para varios procesos relacionados con la respuesta de virus. Esto está de acuerdo con los resultados obtenidos en la figura 5.20, donde se prevé que los procesos de infección viral se activan.

están relacionados con el *p-value* de las categorías y la intensidad del color representa el *z-score* del proceso. La figura 5.20 *arriba*; es la matriz completa con todas las categorías enriquecidas por el análisis de expresión diferencial. Los colores del *z-score* son azules para una disminución prevista, mientras que el color naranja predice una elevación del proceso. Observe que la categoría más naranja corresponde a procesos relacionados con enfermedades infecciosas. La figura 5.20 *abajo*; es un "zoom" en dicha categoría. El único cuadrado azul en este mapa de calor es para la infección bacteriana, los cuadrados grises son infecciones respiratorias, mientras que el resto de los elementos tienen *z-scores* altos. Todos estos procesos están relacionados con respuestas virales. Estos resultados combinados (comunidad OAS2 y mapa de calor) pueden sugerir que en el fenotipo de cáncer de mama HER2+, las células están respondiendo para contrarrestar la enfermedad como en un evento de infección viral. Se necesitan más estudios para aclarar este resultado. Interseantemente, la comunidad OAS2 y las moléculas involucradas en la categoría de infección viral IPA no se comparten. Esto significa que en dos enfoques diferentes, están apareciendo moléculas relacionadas con las respuestas a los virus.

Finalmente, es importante señalar que los principales resultados de este caso de estudio están en proceso de revisión para publicación [311].

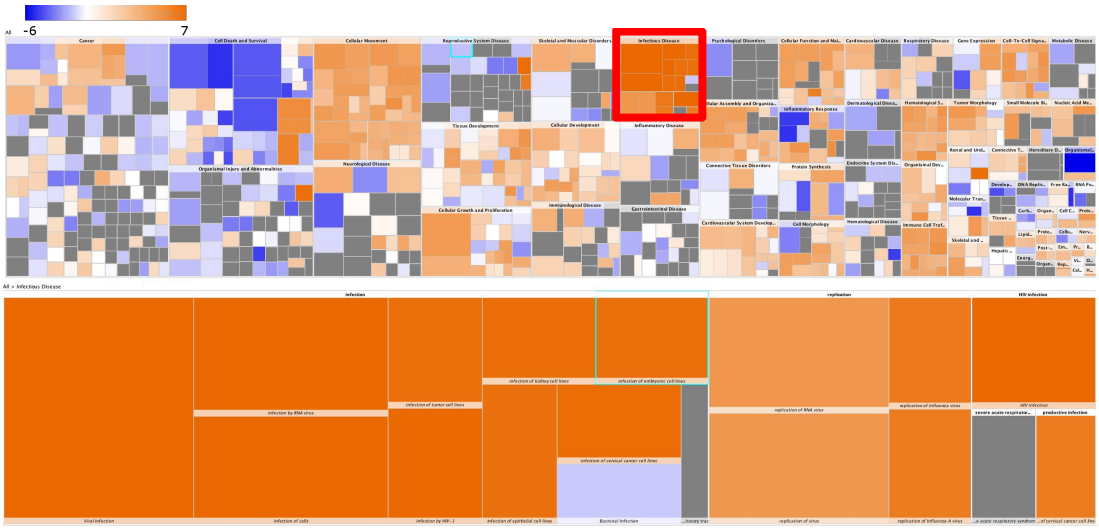


Figura 5.20: **Enfermedades y funciones enriquecidas en el subtipo Her2+ de cáncer de mama**. Esta figura muestra las enfermedades y funciones enriquecidas del subtipo molecular Her2+ de cáncer de mama, de acuerdo con la *Ingenuity Knowledge Base* (IKB). Los tamaños cuadrados son proporcionales al $-log(p_v)$ del conjunto de genes. El código de color representa el *z-score* de las moléculas en cada categoría, dependiendo de la actividad pronosticada que el proceso podría ejercer: los colores azules tienen en cuenta la disminución del proceso, mientras que los colores naranja muestran procesos activados. La mitad superior contiene todas las categorías enriquecidas, separadas por categorías generales. La mitad inferior es un zoom en las categorías "Enfermedades infecciosas" (cuadrado rojo en la mitad superior). Obsérvese como todos los procesos, excepto las enfermedades respiratorias (gris) y las infecciones bacterianas (azul) tienen una activación prevista, lo que significa que las categorías están más activadas que en un control.

Discusión y Conclusiones.

En este trabajo hemos propuesto una metodología computacional matemáticamente formal y robusta, con la cual es posible detectar y estudiar *módulos biológicamente funcionales* en redes de regulación genética de gran escala inferidas a partir de datos experimentales obtenidos de bases de datos publicas. En el caso particular del estudio de las redes de de cáncer, estás fueron inferidas a partir de muestras de biopsias de genoma completo en humano.

Ésta metodología se enmarca en el enfoque actual de la *Biología de Sistemas* que es capaz de integrar grandes datos provenientes de las tecnologías informáticas y *omicas* con las que contamos hoy en día, y con las cuales se puede disponer e integrar una gran cantidad de información biológica. De está manera este enfoque es muy diferente a las *redes de regulación genética booleanas tradicionales* las cuales son insuficientes para modelar redes de genoma completo o de organismos grandes como humano.

Sin duda, el avance de la bioformática y la *Ciencia de datos*, ha tenido una aporte fundamental el estudio de sistemas biológicos, sin embargo este avance no sería posible sin un marco matemático que sustente los análisis y más aún que proporcione no sólo nuevas técnicas sino nuevas formas teóricas de concebir a los sistemas biológicos. Un ejemplo concreto de lo anterior lo proporcionan las redes de co-expresión (como el proporcionando por el método *ARACNE*), que infieren el programa de regulación transcripcional de un cierto fenotipo a través de la *Teoría de la Información* a partir de datos experimentales.

Bajo esta misma filosofía, la teoría de redes complejas o hoy en día llamada *Ciencia de Redes*, proporciona herramientas matemáticas formales para el estudio topologíco (estructura) y de los procesos que pueden ocurrir (dinámica) en una red. Asimismo, la aportación a la biología por parte de la *Ciencia de Redes* es muy importante, dando lugar a nuevas ramas de ambas ciencias como la *Biología de Red* o la *Medicina de Red*, cuyas aplicaciones han sido de gran ayuda a estudiar de en varias escalas problemas de relevancia en las ciencias biológicas. De está manera, en este trabajo en particular analizamos redes de co-expresión de muchos genes ($\sim 10^4$) y muchísimas interacciones entre ellos, las cuales modelamos como Redes Complejas.

Una de las características topológicas más importantes de las *Redes Complejas* es la modularidad de las mismas. Así también, la modularidad es muy importante en general en la biología, pero en particular, como pusimos de manifiesto en en el desarrollo de ésta tesis, las redes biológicas muestran modularidad y una estructura jerárquica intrínseca, la cual es importare estudiar mediante un enfoque formal.

En este trabajo estudiamos la estructura de la comunidad de una red transcripcional, asociada a un factor de transcripción relevante: *MEF2C*. Asimismo, al adoptar un

enfoque de Biología de Sistemas, encontramos la estructura modular para *redes transcripcionales asociadas subtipos moleculares de cáncer de mama* (inferidas a partir de datos experimentales) y realizamos análisis de enriquecimiento funcional para todos los módulos detectados. Finalmente analizamos la *estructura modular jerárquica de la red del subtipo molecular HER2+* para llegar a una descripción más precisa de la función de los procesos asociados a su arquitectura.

Comunidades Funcionales en la red de transcripción de MEF2C.

Encontramos que la estructura modular de la red de *MEF2C* curada a partir de la Base de Datos *Fantom4* contiene información biológicamente significativa. Al analizar el conjunto de genes altamente conectados, se puede observar que estos tienen un papel relevante en la estructura de la red con implicaciones biológicas. Por ejemplo, *GABPA* es un factor de transcripción bien conocido que contiene un motivo de unión al ADN que está involucrado en la activación de la expresión de la *citocromo oxidasa* y el *control nuclear de la función mitocondrial* [320]. Asimismo, *GABPA* pertenece a la familia ETS, y regula diferentes genes blancos asociados con funciones *citoesqueléticas* y control de migración celular [321].

El gen *NFYC*, es un factor de transcripción altamente conservado que es activador de una gran variedad de genes. Está relacionado con la reparación del ADN [322], la regulación de la transcripción [323] y su mal funcionamiento se asocia con el desarrollo de diferentes tipos de carcinomas [324]. Por otro lado, *JUND*, un gen sin intrones, miembro de la familia JUN, y un componente funcional del complejo del factor de transcripción AP1. Se ha propuesto proteger a las células de la senescencias dependientes de p53 y la apoptosis. Este gen es uno de los más importantes relacionados con el cáncer [325, 326, 327, 328, 329, 330, 331, 332, 333], pero también es relevante en eventos apoptóticos y reparación de ADN [334], proliferación [335] y estrés oxidativo [336].

Finalmente, el factor de transcripción *SRF* codifica una proteína nuclear ubicua [337] relacionada con la proliferación celular [338] y diferenciación [339]. Su proteína se une al elemento *SRE* en la región promotora de los genes diana y regula la actividad de muchos genes inmediatos, como *c-fos*, y participa en la regulación del ciclo celular, *apoptosis* y transición epitelio-mesenquima. [340]. Vale la pena mencionar que la comunidad SRF es la más pequeña con procesos enriquecidos de acuerdo con nuestro umbral de criterio ($p_v < 10^{-5}$); sin embargo, los procesos específicos de señalización celular están altamente enriquecidos en esta comunidad. Es obligatorio realizar más investigaciones sobre la funcionalidad de esta comunidad.

Asimismo, el análisis de la estructura modular, a su vez, revela que hay módulos especializados en un procesos particulares (es decir, procesos fuertemente enriquecidos en esas subredes), lo anterior sugiere que la coexpresión de estos conjuntos de genes

(una comunidad) pueden estar directamente asociados con dichos procesos celulares en una forma *adaptativa*. Además, hay varios procesos que están enriquecidos en más de un módulo (ver las columnas correspondientes a la comunidad etiquetada JUND en las figuras 5.3 a 5.6), en este caso es posible argumentar que estos las comunidades colaboran juntas dando *robustez* al sistema en términos de conjuntos de genes.

Un resultado bastante interesante obtenido por el análisis de modularidad, es que en cada una de las comunidades más grandes, el flujo de información está controlado por factores de transcripción (GABPA, NFYC, JUND, etc.), los cuales a su vez, están regulados por MEF2C, lo que refuerza la hipótesis de que MEF2C es un *regulador maestro transcripcional*. Estos últimos resultados nos llevan a argumentar que la estructura modular está fuertemente vinculada a la funcionalidad en la red transcripcional de MEF2C.

Es importante observar que la estructura modular está relacionada con la partición estructural y funcional de la red, en el sentido de que comunidades realizan funciones individuales (como se muestra en las figuras 5.3 - 5.6). Como revela una inspección detallada de la figura 5.2, el flujo de información de regulación entre las comunidades implica patrones de regulación inter-modulares. Por lo tanto, los módulos pueden estar funcionando como *piezas de maquinaria* involucradas en los procesos de varios sub-sistemas que conforman un dispositivo. A nivel local, realizan acciones altamente especializadas que, en el contexto del rendimiento global de la máquina, también colaboran con otros sub-sistemas en el dispositivo.

Comunidades Funcionales en redes de cáncer de mama.

En el caso de las redes transcripcionales asociados a los subtipos moleculares de cáncer de mama, a modo de resumen, presentamos los resultados más relevantes obtenidos con nuestro enfoque:

- La modularidad de cada red de subtipo molecular de cáncer de mama es diferente.
- Existe un patrón único de procesos enriquecidos asociados con módulos para cada red de subtipos.
- A pesar de estas estructuras modulares particulares, el módulo COL5A2_{comm} está presente en todos los subtipos, a pesar de que este módulo es diferente en cada subtipo (es decir, no tiene los mismos genes), comparte procesos relacionados con la matriz extracelular y la formación de fibrillas de colágeno, lo que sugiere robustez en los procesos para el cáncer de mama en general, independientemente de los genes participantes.
- Para la estructura modular del subtipo basal, existen varios procesos únicos que están relacionados con la *apoptosis* y *sistema inmune*.

Como se puede observar a partir del análisis, la estructura de las redes de regulación genética para los subtipos de cáncer de mama está fuertemente asociada con la función. Esta asociación no se puede observar directamente a menos que se analice mediante un enfoque como el proporcionado aquí. Con respecto al subtipo basal, la estrecha corregulación de los genes que participan juntos en los procesos relacionados con la apoptosis puede abrir la posibilidad de explorar terapias dirigidas a esas moléculas o elementos aguas abajo de ellas.

La inmunidad en el subtipo basal no ha sido profundamente explorada. Una metodología como la presentada aquí se puede combinar con diferentes análisis para integrar múltiples fuentes de información en un marco sólido. Un ejemplo de esto se puede encontrar en [341], donde Genome-Wide Association Studies (GWAS) se combinaron con genes reguladores maestros CD4+, para identificar posibles mecanismos de aparición de fenotipos aberrantes.

Pudimos reducir los grados de libertad de nuestros datos. Usando esa información, identificamos estructuras modulares, relacionadas con la manera en que los genes están interconectados. Esa arquitectura no solo es relevante para el análisis topológico, sino también para la funcionalidad biológica. El trabajo adicional incluye posibles aplicaciones en diferentes conjuntos de datos, tecnología RNA-seq, otros cánceres y posibles procedimientos experimentales para investigar la implicación de los procesos relacionados con la apoptosis y la inmunidad en el subtipo basal de cáncer de mama.

En el caso de la arquitectura de la red asociada al subtipo molecular de cáncer de mama HER2+, encontramos que la modularidad refleja asociaciones funcionales con dichos módulos y al analizar su estructura jerárquica se puede obtener una descripción más precisa de la función de los procesos asociados a la arquitectura de la red.

Con nuestro enfoque, fuimos capaces de observar procesos biológicos enriquecidos asociados a pequeños subconjuntos específicos de genes que revelan dos características novedosas principales de este fenotipo patológico: una respuesta viral surgió en el módulo *LCK* y un posible papel de la regulación de *microARN* en el módulo *CNR2*. Además, el análisis de expresión diferencial está ampliamente de acuerdo con la estructura de red modular.

Los resultados presentados aquí son un claro ejemplo de cuán importante es la estructura de red a gran escala para revelar el mecanismo subyacente detrás de una patología particular. La instancia presentada en este trabajo (subtipo de cáncer de mama HER2+) es solo un ejemplo de cuán crucial es el papel del enfoque de ciencias de redes en las ciencias biomédicas.

Conclusiones Generales.

En general, mediante una metodología computacional basada en métodos matemáticamente formales, fuimos capaces de identificar módulos (comunidades) *biológicamente funcionales* en redes de regulación genética. Asimismo, podemos decir que dichos módulos corresponden con funciones biológicas concretas. Lo cual cumple con el objetivo principal planteado en el proyecto doctoral y satisface los objetivos particulares.

Tanto el algoritmo de detección de comunidades como el análisis de enriquecimiento realizados aquí se pueden aplicar a una gran variedad de redes, que se pueden inferir con diferentes enfoques. Esto constituye una característica importante de la metodología presentada aquí. El análisis de enriquecimiento de las comunidades detectadas puede proporcionar información para comprender la complejidad que subyace al proceso de regulación de la transcripción.

De esta manera, hemos puesto de manifiesto que es posible modelar matemáticamente el *programa de regulación transcripcional genético* en humano, en particular en el caso de enfermedades complejas como el cáncer. Sin embargo, es imperativo y necesario mucha más investigación sobre estos temas para comprender mecanismos concretos. No obstante, este trabajo representa mediante un enfoque teórico, un pequeña a aportación para responder a preguntas fundamentales como: ¿de qué manera se regula la expresión de todos genes dentro de una célula, dando paso a un fenotipo particular? y ¿cómo es que estos son capaces de gobernar todas las reacciones que ocurren en la misma?.

La metodología presentada aquí de ninguna manera es concluyente sobre los mecanismos funcionales en cáncer de mama, sin embargo al reducir la dimensionalidad de los datos, es capaz de proporcionar pistas importantes en la dirección de proponer y diseñar experimentos relacionados a nuestros hallazgos. Lo anterior puede lograr eficientar recursos experimentales, técnicos y humanos en la búsqueda del entendimiento del funcionamiento de sistemas biológicos a nivel molecular así como posibles tratamientos farmacológicos para el cáncer de mama.

Con la metodología realizada aquí, pudimos descubrir *módulos funcionales* relacionadas con procesos biológicos específicos que distan mucho de estar elegidos al azar, sino que cuentan con una confianza estadística. Esto es sumamente importante, pues significa que bajo nuestro modelo de red compleja, estamos localizando grupos de genes (dentro de todo un genoma) a partir de las interacciones en la red, es decir, a partir de que estos genes están corregulados de una manera coordinada y que además participan de forma conjunta en proceso biológicos particulares conocidos. Por lo que podemos inferir *sub-unidades regulatorias funcionales* en fenotipos a partir de la estructura de se red de regulación, lo cual puede dirigir futuros estudios experimentales.

Lo cual la muestra una instancia más en la que la estructura de redes de regulación

genética revela funcionalidad. Sin embargo es muy importante señalar que se necesitan experimentos para corroborar los hallazgos presentados aquí, pero vale la pena notar que estos resultados no pudieron ser alcanzados sino solo teniendo en cuenta la estructura de la redes estudiadas.

Finalmente, es valioso mencionar que enfoques como el presentado aquí, son necesarios para reducir dimensionalidad a partir de la gran cantidad de datos omicos con los que se cuenta hoy en día. Asimismo, consideramos que propuestas como la nuestra ayudan a proporcionar elementos necesarios, para generalizar una ***Biología Teórica*** formal a partir de enfoques como los de la *Biología de Sistemas*, la *Ciencia de Redes* y la *Ciencia de Datos*.

Bibliografía

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, p. 561, 1970. 2

[2] F. H. Crick, "On protein synthesis," in *Symp Soc Exp Biol*, vol. 12, p. 8, 1958. 2

[3] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, and et al., "An atlas of combinatorial transcriptional regulation in mouse and man," *Cell*, vol. 140, p. 744,Äì752, Mar 2010. 2

[4] J. Monod and F. Jacob, "General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation," in *Cold Spring Harbor symposia on quantitative biology*, vol. 26, pp. 389–401, Cold Spring Harbor Laboratory Press, 1961. 5

[5] L. Hood and E. V. Rothenberg, "Developmental biologist eric h. davidson, 1937–2015," *Proceedings of the National Academy of Sciences*, vol. 112, no. 44, pp. 13423–13425, 2015. 5

[6] R. J. Britten and E. H. Davidson, "Gene regulation for higher cells: a theory," *Science*, vol. 165, no. 3891, pp. 349–357, 1969. 5

[7] R. J. Britten and E. H. Davidson, "Repetitive and non-repetitive dna sequences and a speculation on the origins of evolutionary novelty," *Quarterly Review of Biology*, pp. 111–138, 1971. 5, 7

[8] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, *et al.*, "A genomic regulatory network for development," *science*, vol. 295, no. 5560, pp. 1669–1678, 2002. 6

[9] E. Sodergren, G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, R. C. Angerer, L. M. Angerer, M. I. Arnone, D. R. Burgess, R. D. Burke, *et al.*, "The genome of the sea urchin strongylocentrotus purpuratus," *Science*, vol. 314, no. 5801, pp. 941–952, 2006. 6

[10] E. Davidson, "Special issue: The sea urchin genome," *Developmental Biology*, vol. 300, no. 1, p. 1, 2006. 6

[11] E. H. Davidson and M. S. Levine, "Properties of developmental gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 105, no. 51, pp. 20063–20066, 2008. 6, 7

[12] I. S. Peter, E. Faure, and E. H. Davidson, "Predictive computation of genomic logic processing functions in embryonic development," *Proceedings of the National Academy of Sciences*, vol. 109, no. 41, pp. 16434–16442, 2012. 7

[13] E. H. Davidson and D. H. Erwin, "Gene regulatory networks and the evolution of animal body plans," *Science*, vol. 311, no. 5762, pp. 796–800, 2006. 7

[14] S. Istrail and E. H. Davidson, "Logic functions of the genomic cis-regulatory code," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4954–4959, 2005. 7

[15] E. Davidson and M. Levin, "Gene regulatory networks," *Proceedings of the national academy of sciences of the United States of America*, vol. 102, no. 14, pp. 4935–4935, 2005. 7

[16] S. S. Damle and E. H. Davidson, "Synthetic in vivo validation of gene network circuitry," *Proceedings of the National Academy of Sciences*, vol. 109, no. 5, pp. 1548–1553, 2012. 7

[17] E. H. Davidson, "Emerging properties of animal gene regulatory networks," *Nature*, vol. 468, no. 7326, pp. 911–920, 2010. 7

[18] M. Levine and E. H. Davidson, "Gene regulatory networks for development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4936–4942, 2005. 8

[19] V. F. Hinman and E. H. Davidson, "Evolutionary plasticity of developmental gene regulatory network architecture," *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19404–19409, 2007. 8

[20] D. H. Erwin and E. H. Davidson, "The evolution of hierarchical gene regulatory networks," *Nature Reviews Genetics*, vol. 10, no. 2, pp. 141–148, 2009. 8

[21] M. Sugita, "Functional analysis of chemical systems in vivo using a logical circuit equivalent. ii. the idea of a molecular automaton," *Journal of Theoretical Biology*, vol. 4, no. 2, pp. 179–192, 1963. 8

[22] R. Thomas, "Boolean formalization of genetic control circuits," *Journal of theoretical biology*, vol. 42, no. 3, pp. 563–585, 1973. 8

[23] S. A. Kauffman, "Sequential DNA replication and the control of differences in gene activity between sister chromatids a possible factor in cell differentiation," *Journal of theoretical biology*, vol. 17, no. 3, pp. 483–497, 1967. 8, 10

[24] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of theoretical biology*, vol. 22, no. 3, pp. 437–467, 1969. 8, 10

[25] S. Wolfram, "Statistical mechanics of cellular automata," *Reviews of modern physics*, vol. 55, no. 3, p. 601, 1983. 9

[26] S. Kauffman, "Gene regulation networks: A theory for their global structure and behaviors," in *Current topics in developmental biology*, vol. 6, pp. 145–182, Elsevier, 1971. 10

[27] L. Glass and S. A. Kauffman, "The logical analysis of continuous, non-linear biochemical control networks," *Journal of theoretical Biology*, vol. 39, no. 1, pp. 103–129, 1973. 10, 11

[28] S. A. Kauffman, *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA, 1993. 10

[29] B. Derrida and Y. Pomeau, "Random networks of automata: a simple annealed approximation," *EPL (Europhysics Letters)*, vol. 1, no. 2, p. 45, 1986. 10

[30] R. V. Solé, P. Fernández, and S. A. Kauffman, "Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion," *arXiv preprint q-bio/0311013*, 2003. 10

[31] R. Albert, "Boolean modelingof genetic regulatory networks," in *Complex networks*, pp. 459–481, Springer, 2004. 10

[32] M. Aldana, E. Balleza, S. Kauffman, and O. Resendiz, "Robustness and evolvability in genetic regulatory networks," *Journal of theoretical biology*, vol. 245, no. 3, pp. 433–448, 2007. 10

[33] M. Davidich and S. Bornholdt, "The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model," *Journal of Theoretical Biology*, vol. 255, no. 3, pp. 269–277, 2008. 10

[34] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, *et al.*, "Transcriptional regulatory networks in saccharomyces cerevisiae," *science*, vol. 298, no. 5594, pp. 799–804, 2002. 10

[35] J. Espinal, M. Aldana, A. Guerrero, C. Wood, A. Darszon, and G. Martínez-Mekler, "Discrete dynamics model for the speract-activated ca2+ signaling network relevant to sperm motility," *PloS one*, vol. 6, no. 8, p. e22619, 2011. 10

[36] J. Espinal-Enríquez, A. Darszon, A. Guerrero, and G. Martínez-Mekler, "In silico determination of the effect of multi-target drugs on calcium dynamics signaling network underlying sea urchin spermatozoa motility," *PloS one*, vol. 9, no. 8, p. e104451, 2014. 10

[37] J. Espinal-Enríquez, D. A. Priego-Espinosa, A. Darszon, C. Beltrán, and G. Martínez-Mekler, "Network model predicts that catsper is the main ca 2+ channel in the regulation of sea urchin sperm motility," *Scientific Reports*, vol. 7, no. 1, p. 4236, 2017. 10

[38] L. Sánchez and D. Thieffry, "A logical analysis of the Drosophila gap-gene system," *Journal of theoretical Biology*, vol. 211, no. 2, pp. 115–141, 2001. 10

[39] A. Ghysen and R. Thomas, "The formation of sense organs in Drosophila: a logical approach," *BioEssays*, vol. 25, no. 8, pp. 802–807, 2003. 10

[40] T. Koide, T. Hayata, and K. W. Cho, "Xenopus as a model system to study transcriptional regulatory networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4943–4948, 2005. 10

[41] L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla, "Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis.," *Bioinformatics*, vol. 15, no. 7, pp. 593–606, 1999. 11

[42] E. R. Alvarez-Buylla, Á. Chaos, M. Aldana, M. Benítez, Y. Cortes-Poza, C. Espinosa-Soto, D. A. Hartasánchez, R. B. Lotto, D. Malkin, G. J. E. Santos, *et al.*, "Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape," *Plos one*, vol. 3, no. 11, p. e3626, 2008. 11

[43] E. R. Jackson, D. Johnson, and W. G. Nash, "Gene networks in development," *Journal of theoretical Biology*, vol. 119, no. 4, pp. 379–396, 1986. 11

[44] S. A. Teichmann and M. M. Babu, "Gene regulatory network growth by duplication," *Nature genetics*, vol. 36, no. 5, pp. 492–496, 2004. 11

[45] K. Voordeckers, K. Pougach, and K. J. Verstrepen, "How do regulatory networks evolve and expand throughout evolution?," *Current opinion in biotechnology*, vol. 34, pp. 180–188, 2015. 11

[46] C. A. Grove and A. J. Walhout, "Transcription factor functionality and transcription regulatory networks," *Molecular BioSystems*, vol. 4, no. 4, pp. 309–314, 2008. 11

[47] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008. 11

[48] H. Singh, K. L. Medina, and J. M. Pongubala, "Contingent gene regulatory networks and B cell fate specification," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 4949–4953, 2005. 11

[49] R. V. Solé, I. Salazar-Ciudad, and J. Garcia-Fernández, "Common pattern formation, modularity and phase transitions in a gene network model of morphogenesis," *Physica A: Statistical Mechanics and its Applications*, vol. 305, no. 3, pp. 640–654, 2002. 11

[50] S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber, "Cell fates as high-dimensional attractor states of a complex gene regulatory network," *Physical review letters*, vol. 94, no. 12, p. 128701, 2005. 11

[51] R. Edwards and L. Glass, "A calculus for relating the dynamics and structure of complex biological networks," *Adventures in Chemical Physics: A Special Volume of Advances in Chemical Physics*, vol. 132, pp. 151–178, 2005. 11

[52] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002. 11

[53] A. Ribeiro, R. Zhu, and S. A. Kauffman, "A general modeling strategy for gene regulatory networks with stochastic dynamics," *Journal of Computational Biology*, vol. 13, no. 9, pp. 1630–1639, 2006. 11

[54] V. Sevim and P. A. Rikvold, "Chaotic gene regulatory networks can be robust against mutations and noise," *Journal of theoretical biology*, vol. 253, no. 2, pp. 323–332, 2008. 11

[55] I. H. G. S. Consortium *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, p. 860, 2001. 13

[56] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, R. A. Young, *et al.*, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.," in *Pacific symposium on biocomputing*, vol. 6, p. 266, 2001. 18

[57] X. Zhou, X. Wang, and E. R. Dougherty, "Construction of genomic networks using mutual-information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design," *Signal Processing*, vol. 83, no. 4, pp. 745–761, 2003. 18

[58] B. Zhang, S. Horvath, *et al.*, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, p. 1128, 2005. 18, 19, 51

[59] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data," *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2005. 18

[60] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC bioinformatics*, vol. 7, no. Suppl 1, p. S7, 2006. 18, 160

[61] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948. 19, 117, 132, 146

[62] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. No. v. 1 in The Mathematical Theory of Communication, University of Illinois Press, 1949. 19, 117, 132, 146

[63] W. Weaver, "Recent contributions to the mathematical theory of communication," *ETC: a review of general semantics*, pp. 261–281, 1953. 19, 117, 132, 146

[64] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. No. pt. 11 in Illini books, University of Illinois Press, 1963. 19, 117, 132, 146

[65] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001. 19, 117, 132, 146

[66] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 19, 132

[67] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the national academy of sciences*, vol. 107, no. 14, pp. 6286–6291, 2010. 19

[68] F. Emmert-Streib, R. de Matos Simoes, P. Mullan, B. Haibe-Kains, and M. Dehmer, "The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks," *Frontiers in genetics*, vol. 5, p. 15, 2014. 19

[69] G. de Anda-Jáuregui, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus, "Transcriptional network architecture of breast cancer molecular subtypes," *Frontiers in Physiology*, vol. 7, 2016. 19, 34, 158, 160, 174, 175

[70] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004. 19, 40, 45, 51

[71] M. Á. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the national academy of sciences*, vol. 106, no. 16, pp. 6483–6488, 2009. 19, 40, 45, 53

[72] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000. 20

[73] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002. 21, 51, 84

[74] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004. 22, 164

[75] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007. 25

[76] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011. 25

[77] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, 2011. 26

[78] M. Gustafsson, C. E. Nestor, H. Zhang, A.-L. Barabási, S. Baranzini, S. Brunak, K. F. Chung, H. J. Federoff, A.-C. Gavin, R. R. Meehan, *et al.*, "Modules, networks and systems medicine for understanding disease and aiding diagnosis," *Genome medicine*, vol. 6, no. 10, p. 82, 2014. 26

[79] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature communications*, vol. 5, p. 4212, 2014. 27

[80] S. D. Ghiassian, J. Menche, and A.-L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome," *PLoS Comput Biol*, vol. 11, no. 4, p. e1004120, 2015. 27, 84, 160

[81] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, 2015. 28, 84, 160

[82] A. Sharma, J. Menche, C. C. Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, *et al.*, "A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma," *Human molecular genetics*, vol. 24, no. 11, pp. 3005–3020, 2015. 28

[83] S. D. Ghiassian, J. Menche, D. I. Chasman, F. Giulianini, R. Wang, P. Ricchiuto, M. Aikawa, H. Iwata, C. Müller, T. Zeller, *et al.*, "Endophenotype Network Models: Common Core of Complex Diseases," *Scientific reports*, vol. 6, p. 27414, 2016. 29

[84] M. Kitsak, A. Sharma, J. Menche, E. Guney, S. D. Ghiassian, J. Loscalzo, and A.-L. Barabási, "Tissue specificity of human disease module," *Scientific reports*, vol. 6, p. 35241, 2016. 30

[85] A. Vinayagam, T. E. Gibson, H.-J. Lee, B. Yilmazel, C. Roesel, Y. Hu, Y. Kwon, A. Sharma, Y.-Y. Liu, N. Perrimon, *et al.*, "Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets," *Proceedings of the National Academy of Sciences*, p. 201603992, 2016. 30

[86] J. Menche, E. Guney, A. Sharma, P. J. Branigan, M. J. Loza, F. Baribaud, R. Dobrin, and A.-L. Barabási, "Integrating personalized gene expression profiles into predictive disease-associated gene pools," *NPJ systems biology and applications*, vol. 3, no. 1, p. 10, 2017. 31

[87] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011. 32, 176

[88] J. West, G. Bianconi, S. Severini, and A. E. Teschendorff, "Differential network entropy reveals cancer system hallmarks," *Scientific reports*, vol. 2, 2012. 32, 74

[89] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002. 37, 38, 44, 47, 48, 49, 51, 64, 65, 69, 70, 71, 72, 74

[90] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003. 37, 38, 44, 47, 48, 49, 51, 53, 59, 60, 64, 65, 66, 69, 70, 71, 72, 74

[91] M. Newman, *Networks: an introduction*. Oxford university press, 2010. 37, 41, 42, 46, 48, 51, 60, 72, 74

[92] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances in physics*, vol. 51, no. 4, pp. 1079–1187, 2002. 37, 38, 47, 49, 51, 64, 65, 69, 71, 72, 74

[93] M. Newman, A.-L. Barabasi, and D. J. Watts, *The structure and dynamics of networks*. Princeton University Press, 2011. 37, 51, 60, 72, 74

[94] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006. 37, 38, 47, 48, 49, 51, 53, 60, 64, 65, 69, 72, 74

[95] G. Caldarelli and A. Vespignani, *Large scale structure and dynamics of complex networks*. World Scientific, 2007. 37, 38, 47, 49, 51, 64, 65, 69, 72, 74

[96] M. Aldana, "Redes complejas," 2006. 37, 51, 61, 63, 64

[97] M. Aldana, "Redes complejas: Estructura, dinámica y evolución," *Recuperado de http://www. fis. unam. mx/˜ max/MyWebPage/notastwocolumn. pdf Alvarado Prada, LE (2008). Investigación colectiva: aproximaciones teórico-metodológicas. Estudios Pedagógicos*, vol. 34, no. 1, pp. 157–172, 2011. 37, 51, 61, 63, 64

[98] C. Berge, *The theory of graphs*. Courier Corporation, 1962. 37

[99] F. Harary, "Graph theory," *Reading, Addison Wesley*, 1969. 37

[100] B. Bollobás, *Modern graph theory*, vol. 184. Springer Science & Business Media, 2013. 37, 43

[101] P. Erdös and A. Rényi, "On random graphs, I," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959. 37, 51, 61, 158

[102] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 17-61, p. 43, 1960. 37, 51, 61

[103] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001. 38

[104] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, vol. 75, no. 2, p. 027105, 2007. 40, 45

[105] M. Barahona and L. M. Pecora, "Synchronization in small-world systems," *Physical review letters*, vol. 89, no. 5, p. 054101, 2002. 43

[106] T. Nishikawa, A. E. Motter, Y.-C. Lai, and F. C. Hoppensteadt, "Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize?," *Physical review letters*, vol. 91, no. 1, p. 014101, 2003. 43

[107] F. R. Chung, *Spectral graph theory*. No. 92, American Mathematical Soc., 1997. 43

[108] A. Pothen, "Graph partitioning algorithms with applications to scientific computing," in *Parallel Numerical Algorithms*, pp. 323–368, Springer, 1997. 43, 44

[109] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak mathematical journal*, vol. 23, no. 2, pp. 298–305, 1973. 44, 106

[110] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Czechoslovak Mathematical Journal*, vol. 25, no. 4, pp. 619–633, 1975. 44, 106

[111] G. Bianconi and A.-L. Barabási, "Bose-Einstein condensation in complex networks," *Physical review letters*, vol. 86, no. 24, p. 5632, 2001. 44

[112] R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera, *Statistical mechanics of complex networks*, vol. 625. Springer Science & Business Media, 2003. 44

[113] D. J. Watts, *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 1999. 47, 48

[114] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967. 48, 60, 66

[115] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks.," *nature*, vol. 393, no. 6684, pp. 440–442, 1998. 49, 51, 66, 68, 104

[116] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, "Generalizations of the clustering coefficient to weighted complex networks," *Physical Review E*, vol. 75, no. 2, p. 027105, 2007. 51

[117] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge University Press, 2008. 51, 72, 74

[118] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198701, 2001. 54

[119] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953. 56

[120] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998. 56, 118, 167, 187

[121] M. E. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003. 59, 60

[122] R. Solomonoff and A. Rapoport, "Connectivity of random nets," *Bulletin of Mathematical Biology*, vol. 13, no. 2, pp. 107–117, 1951. 61

[123] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009. 69

[124] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999. 69, 70

[125] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001. 69

[126] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, *et al.*, "The degree sequence of a scale-free random graph process," *Random Structures & Algorithms*, vol. 18, no. 3, pp. 279–290, 2001. 71

[127] G. Bianconi, "Entropy of network ensembles," *Physical Review E*, vol. 79, no. 3, p. 036114, 2009. 74

[128] K. Anand and G. Bianconi, "Entropy measures for networks: Toward an information theory of complex topologies," *Physical Review E*, vol. 80, no. 4, p. 045102, 2009. 74

[129] K.-i. Hashimoto, "Zeta functions of finite graphs and representations of p-adic groups," *Automorphic forms and geometry of arithmetic varieties.*, pp. 211–280, 1989. 74, 108

[130] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20935–20940, 2013. 74, 108

[131] O. Angel, J. Friedman, and S. Hoory, "The non-backtracking spectrum of the universal cover of a graph," *Transactions of the American Mathematical Society*, vol. 367, no. 6, pp. 4287–4318, 2015. 74

[132] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, 2015. 74

[133] I. A. Kovács and A.-L. Barabási, "Network science: Destruction perfected," *Nature*, vol. 524, no. 7563, pp. 38–39, 2015. 74

[134] X. Teng, S. Pei, F. Morone, and H. A. Makse, "Collective Influence of Multiple Spreaders Evaluated by Tracing Real Information Flow in Large-Scale Social Networks," *arXiv preprint arXiv:1606.02740*, 2016. 75

[135] B. Karrer, M. E. Newman, and L. Zdeborová, "Percolation on sparse networks," *Physical review letters*, vol. 113, no. 20, p. 208702, 2014. 75

[136] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, "Mathematical formulation of multilayer networks," *Physical Review X*, vol. 3, no. 4, p. 041022, 2013. 75

[137] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014. 75

[138] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014. 75

[139] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, "Structural reducibility of multilayer networks," *Nature communications*, vol. 6, 2015. 76

[140] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, "The physics of multilayer networks," *arXiv preprint arXiv:1604.02021*, 2016. 76

[141] M. De Domenico, M. A. Porter, and A. Arenas, "MuxViz: a tool for multilayer analysis and visualization of networks," *Journal of Complex Networks*, p. cnu038, 2014. 76

[142] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, "Ranking in interconnected multilayer networks reveals versatile nodes," *Nature communications*, vol. 6, 2015. 76

[143] G. Bianconi, "Statistical mechanics of multiplex networks: Entropy and overlap," *Physical Review E*, vol. 87, no. 6, p. 062806, 2013. 76

[144] G. Bianconi and S. N. Dorogovtsev, "Multiple percolation transitions in a configuration model of a network of networks," *Physical Review E*, vol. 89, no. 6, p. 062814, 2014. 76

[145] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks," *Physical Review E*, vol. 70, no. 2, p. 025101, 2004. 77, 88, 95, 110, 112, 136, 149

[146] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010. 77, 79, 90, 100, 104, 115, 123, 127, 134

[147] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016. 77, 90, 100

[148] J. Moody, "Race, school integration, and friendship segregation in america," *American journal of Sociology*, vol. 107, no. 3, pp. 679–716, 2001. 79

[149] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977. 80, 123, 124

[150] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 270, no. Suppl 2, pp. S186–S188, 2003. 81, 123, 126

[151] M. A. Porter, P. J. Mucha, M. E. Newman, and C. M. Warmbrand, "A network analysis of committees in the us house of representatives," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 20, pp. 7057–7062, 2005. 81

[152] M. A. Porter, P. J. Mucha, M. E. Newman, and A. J. Friend, "Community structure in the united states house of representatives," *Physica A: Statistical Mechanics and its Applications*, vol. 386, no. 1, pp. 414–438, 2007. 81

[153] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, "Community structure in congressional cosponsorship networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 7, pp. 1705–1712, 2008. 81

[154] J. Espinal-Enriquez, J. M. Siqueiros-García, R. García-Herrera, and S. A. Alcalá-Corona, "A literature-based approach to a narco-network," in *International Conference on Social Informatics*, pp. 97–101, Springer, 2014. 81

[155] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404–409, 2001. 81

[156] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008. 82, 89, 96, 117, 118, 138, 139, 140, 144, 167

[157] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002. 82, 86, 88, 89, 103, 104, 109, 123, 125, 127, 128, 131

[158] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in complex systems*, vol. 6, no. 04, pp. 565–573, 2003. 83

[159] J. Reichardt and S. Bornholdt, "Partitioning and modularity of graphs with arbitrary degree distribution," *Physical Review E*, vol. 76, no. 1, p. 015102, 2007. 83, 112

[160] J. Reichardt and S. Bornholdt, "Detecting fuzzy community structures in complex networks with a Potts model," *Physical Review Letters*, vol. 93, no. 21, p. 218701, 2004. 83, 96, 113

[161] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006. 83, 87, 96, 113

[162] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008. 83, 110, 118, 138, 139, 149

[163] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012. 84

[164] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006. 84, 111

[165] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006. 84, 111

[166] R. V. Solé and S. Valverde, "Spontaneous emergence of modularity in cellular networks," *Journal of The Royal Society Interface*, vol. 5, no. 18, pp. 129–133, 2008. 84

[167] R. V. Solé, S. Valverde, and C. Rodriguez-Caso, "Modularity in biological networks," *Biological Networks. World Scientific, Singapore*, pp. 21–40, 2006. 84

[168] S. Hüffner, *Modularity in Biological Networks*. PhD thesis, Freie Universität Berlin Berlin, 2014. 84

[169] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007. 84

[170] L. Cantini, E. Medico, S. Fortunato, and M. Caselle, "Detection of gene communities in multi-networks reveals cancer drivers," *Scientific reports*, vol. 5, 2015. 84, 90, 160

[171] G. Didier, C. Brun, and A. Baudot, "Identifying communities from multiplex biological networks," *PeerJ*, vol. 3, p. e1525, 2015. 84

[172] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi, "Network-based in silico drug efficacy screening," *Nature communications*, vol. 7, 2016. 84

[173] A. Zhang, *Protein interaction networks: computational analysis*. Cambridge University Press, 2009. 85

[174] P. F. Jonsson, T. Cavanna, D. Zicha, and P. A. Bates, "Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis," *BMC bioinformatics*, vol. 7, no. 1, p. 2, 2006. 85

[175] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005. 85, 90, 122, 138

[176] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, *et al.*, "Evidence for dynamically organized modularity in the yeast protein–protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004. 85

[177] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 1128–1133, 2003. 85, 89

[178] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 21, pp. 12123–12128, 2003. 85

[179] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006. 86

[180] R. Dunn, F. Dudbridge, and C. M. Sanderson, "The use of edge-betweenness clustering to investigate biological function in protein interaction networks," *BMC bioinformatics*, vol. 6, no. 1, p. 39, 2005. 86

[181] V. Farutin, K. Robison, E. Lightcap, V. Dancik, A. Ruttenberg, S. Letovsky, and J. Pradines, "Edge-count probabilities for the identification of local protein communities and their organization," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 3, pp. 800–818, 2006. 87

[182] T. Z. Sen, A. Kloczkowski, and R. L. Jernigan, "Functional clustering of yeast proteins from the protein-protein interaction network," *BMC bioinformatics*, vol. 7, no. 1, p. 355, 2006. 87

[183] A. C. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, "The function of communities in protein interaction networks at multiple scales," *BMC systems biology*, vol. 4, no. 1, p. 100, 2010. 87

[184] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000. 87, 89, 152, 187

[185] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, no. 7028, pp. 895–900, 2005. 88, 110

[186] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003. 88

[187] M. Zhan, "Deciphering modular and dynamic behaviors of transcriptional networks," *Genomic medicine*, vol. 1, no. 1-2, pp. 19–28, 2007. 88

[188] J. Zhang and S. Zhang, "Modular Organization of Gene Regulatory Networks," *Encyclopedia of Systems Biology*, pp. 1437–1441, 2013. 88

[189] C. Cheng, E. Andrews, K.-K. Yan, M. Ung, D. Wang, and M. Gerstein, "An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome," *Genome biology*, vol. 16, no. 1, p. 1, 2015. 88

[190] H. Xu and S. Wang, "Research on functional modules of gene regulatory network," in *Advancing Computing, Communication, Control and Management*, pp. 264–271, Springer, 2010. 88

[191] A. Gyorgy and D. Del Vecchio, "Modular composition of gene transcription networks," *PLoS Comput Biol*, vol. 10, no. 3, p. e1003486, 2014. 88

[192] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002. 88

[193] H. Mohamadlou, G. J. Podgorski, and N. S. Flann, "Motifs Within Genetic Regulatory Networks Increase Organization During Pattern Formation," in *International Conference on Information Processing in Cells and Tissues*, pp. 103–113, Springer, 2015. 88

[194] Y. Qi and H. Ge, "Modularity and dynamics of cellular networks," *PLoS Comput Biol*, vol. 2, no. 12, p. e174, 2006. 88

[195] D. M. Wilkinson and B. A. Huberman, "A method for finding communities of related genes," *proceedings of the national Academy of sciences*, vol. 101, no. suppl 1, pp. 5241–5248, 2004. 88, 90, 104, 160

[196] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," in *Communities and technologies*, pp. 81–96, Springer, 2003. 88, 104

[197] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *science*, vol. 302, no. 5643, pp. 249–255, 2003. 88

[198] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003. 88

[199] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2981–2986, 2004. 88

[200] A. S. Cristino, R. F. Andrade, and L. da Fontoura Costa, "Detecting and Characterizing the Modular Structure of the Yeast Transcription Network," in *Complex Networks*, pp. 35–46, Springer, 2009. 88

[201] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, *et al.*, "Computational discovery of gene modules and regulatory networks," *Nature biotechnology*, vol. 21, no. 11, pp. 1337–1342, 2003. 88

[202] J. V. Oliveira, A. F. de Brito, C. T. Braconi, C. C. de Melo Freire, A. Iamarino, and P. M. de Andrade Zanotto, "Modularity and evolutionary constraints in a baculovirus gene regulatory network," *BMC systems biology*, vol. 7, no. 1, p. 1, 2013. 89

[203] J. A. Freyre-González, A. M. Manjarrez-Casas, E. Merino, M. Martinez-Nuñez, E. Perez-Rueda, and R.-M. Gutiérrez-Ríos, "Lessons from the modular organization of the transcriptional regulatory network of bacillus subtilis," *BMC systems biology*, vol. 7, no. 1, p. 1, 2013. 89

[204] O. Resendis-Antonio, J. A. Freyre-Gonzalez, R. Menchaca-Mendez, R. M. Gutiérrez-Ríos, A. Martínez-Antonio, C. Avila-Sanchez, and J. Collado-Vides, "Modular analysis of the transcriptional regulatory network of e. coli," *TRENDS in Genetics*, vol. 21, no. 1, pp. 16–20, 2005. 89

[205] J. Sanz, E. Cozzo, J. Borge-Holthoefer, and Y. Moreno, "Topological effects of data incompleteness of gene regulatory networks," *BMC systems biology*, vol. 6, no. 1, p. 110, 2012. 89

[206] R. Chauhan, J. Ravi, P. Datta, T. Chen, D. Schnappinger, K. E. Bassler, G. Balázsi, and M. L. Gennaro, "Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis," *Nature communications*, vol. 7, 2016. 89

[207] M. J. Barber, "Modularity and community detection in bipartite networks," *Physical Review E*, vol. 76, no. 6, p. 066102, 2007. 89

[208] L. Mao, J. L. Van Hemert, S. Dash, and J. A. Dickerson, "Arabidopsis gene co-expression network and its functional modules," *BMC bioinformatics*, vol. 10, no. 1, p. 1, 2009. 89

[209] S. M. Van Dongen, *Graph clustering by flow simulation*. PhD thesis, 2001. 89, 116, 134

[210] S. A. Alcalá-Corona, T. E. Velázquez-Caldelas, J. Espinal-Enríquez, and E. Hernández-Lemus, "Community structure reveals biologically functional modules in MEF2C transcriptional regulatory network," *Frontiers in physiology*, vol. 7, 2016. 89, 158, 160, 164, 173

[211] H. Suzuki, A. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. de Hoon, *et al.*, "The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line," *Nature genetics*, vol. 41, no. 5, pp. 553–562, 2009. 89, 156

[212] A. Srivastava, S. Kumar, and R. Ramaswamy, "Two-layer modular analysis of gene and protein networks in breast cancer," *BMC systems biology*, vol. 8, no. 1, p. 1, 2014. 89

[213] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004. 89, 94, 109, 110

[214] Z. Shi, C. K. Derow, and B. Zhang, "Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression," *BMC systems biology*, vol. 4, no. 1, p. 1, 2010. 89

[215] S. A. Alcalá-Corona, G. De Anda Jáuregui, J. Espinal-Enríquez, and E. H.-L. Hernández-Lemus, "Network modularity in breast cancer molecular subtypes," *Frontiers in Physiology*, vol. 8, p. 915, 2017. 89, 160, 164, 185

[216] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, "Characterizing the community structure of complex networks," *PloS one*, vol. 5, no. 8, p. e11976, 2010. 90

[217] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *science*, vol. 328, no. 5980, pp. 876–878, 2010. 90

[218] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010. 90

[219] N. Gulbahce and S. Lehmann, "The art of community detection," *BioEssays*, vol. 30, no. 10, pp. 934–938, 2008. 90

[220] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *Acm computing surveys (csur)*, vol. 45, no. 4, p. 43, 2013. 90

[221] B. Tang, H.-K. Hsu, P.-Y. Hsu, R. Bonneville, S.-S. Chen, T. H. Huang, and V. X. Jin, "Hierarchical modularity in ERα transcriptional network is associated with distinct functions and implicates clinical outcomes," *Scientific reports*, vol. 2, 2012. 90

[222] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature methods*, vol. 9, no. 8, pp. 796–804, 2012. 90, 160

[223] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature genetics*, vol. 40, no. 7, pp. 854–861, 2008. 90, 160

[224] M. E. Newman and M. Girvan, "Mixing patterns and community structure in networks," in *Statistical mechanics of complex networks*, pp. 66–87, Springer, 2003. 90

[225] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004. 92, 93, 103, 109

[226] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007. 94, 95, 108, 112

[227] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009. 94

[228] M. E. Newman, "Analysis of weighted networks," *Physical review E*, vol. 70, no. 5, p. 056131, 2004. 95

[229] C. P. Massen and J. P. Doye, "Thermodynamics of community structure," *arXiv preprint cond-mat/0610077*, 2006. 95, 110

[230] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity," *New Journal of Physics*, vol. 9, no. 6, p. 176, 2007. 95

[231] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004. 100, 101, 104

[232] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967. 102

[233] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004. 104

[234] S. Fortunato, V. Latora, and M. Marchiori, "Method to find community structures based on information centrality," *Physical review E*, vol. 70, no. 5, p. 056104, 2004. 105

[235] L. Donetti and M. A. Muñoz, "Detecting network communities: a new systematic and efficient algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, no. 10, p. P10012, 2004. 106, 139

[236] A. Capocci, V. D. Servedio, G. Caldarelli, and F. Colaiori, "Detecting communities in large networks," *Physica A: Statistical Mechanics and its Applications*, vol. 352, no. 2, pp. 669–676, 2005. 107

[237] M. Newman, "Spectral community detection in sparse networks," *arXiv preprint arXiv:1308.6494*, 2013. 108

[238] A. Singh and M. D. Humphries, "Finding communities in sparse networks," *Scientific reports*, vol. 5, p. 8828, 2015. 108

[239] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On finding graph clusterings with maximum modularity," in *International Workshop on Graph-Theoretic Concepts in Computer Science*, pp. 121–132, Springer, 2007. 109

[240] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983. 110, 149

[241] X. Zhang and M. Newman, "Multiway spectral community detection in networks," *Physical Review E*, vol. 92, no. 5, p. 052808, 2015. 112

[242] J. Reichardt and S. Bornholdt, "When are networks truly modular?," *Physica D: Nonlinear Phenomena*, vol. 224, no. 1-2, pp. 20–26, 2006. 112

[243] A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," *Physical review E*, vol. 84, no. 6, p. 066122, 2011. 112

[244] P. Ronhovde and Z. Nussinov, "Local resolution-limit-free potts model for community detection," *Physical Review E*, vol. 81, no. 4, p. 046114, 2010. 114, 138, 139

[245] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007. 114

[246] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, "Synchronization reveals topological scales in complex networks," *Physical review letters*, vol. 96, no. 11, p. 114102, 2006. 115

[247] Y. Kuramoto, *Chemical oscillations, waves, and turbulence*, vol. 19. Springer Science & Business Media, 2012. 115

[248] H. Zhou, "Network landscape from a brownian particle's perspective," *Physical Review E*, vol. 67, no. 4, p. 041908, 2003. 115

[249] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *International Symposium on Computer and Information Sciences*, pp. 284–293, Springer, 2005. 115

[250] S. A. Alcalá-Corona and y. Padilla, Pablo, "A diffusion-based algorithm for community detection in networks," 116

[251] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952. 117, 144, 145

[252] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009. 118

[253] M. Rosvall and C. T. Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems," *PloS one*, vol. 6, no. 4, p. e18209, 2011. 118, 150

[254] A. V. Esquivel and M. Rosvall, "Compression of flow can reveal overlapping-module organization in networks," *Physical Review X*, vol. 1, no. 2, p. 021025, 2011. 118

[255] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PloS one*, vol. 6, no. 4, p. e18961, 2011. 120, 138, 139

[256] B. Bollobás, "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs," *European Journal of Combinatorics*, vol. 1, no. 4, pp. 311–316, 1980. 120

[257] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995. 120

[258] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, p. 016107, 2011. 120

[259] T. P. Peixoto, "Nonparametric weighted stochastic block models," *Physical Review E*, vol. 97, no. 1, p. 012306, 2018. 121

[260] M. E. Newman, "Spectral methods for community detection and graph partitioning," *Physical Review E*, vol. 88, no. 4, p. 042822, 2013. 122

[261] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures and Algorithms*, vol. 18, no. 2, pp. 116–140, 2001. 127

[262] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E*, vol. 78, no. 4, p. 046110, 2008. 127, 129, 136

[263] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, no. 1, p. 016118, 2009. 131, 136

[264] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009. 131, 134, 136

[265] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005. 132, 134

[266] S. A. Alcalá-Corona, S. Sandoval-Motta, J. Espinal-Enríquez, and E. H.-L. Hernández-Lemus, "Modularity in biological networks: State of the art," 2018. 140

[267] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W741–W748, 2005. 152

[268] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "Web-based gene set analysis toolkit (webgestalt): update 2013," *Nucleic acids research*, vol. 41, no. W1, pp. W77–W83, 2013. 152

[269] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, "Pathway analysis: State of the art," *Frontiers in physiology*, vol. 6, 2015. 152

[270] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005. 152

[271] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a go category within a class of genes: which test?," *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2006. 153

[272] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995. 154

[273] K. Baca-López, M. Mayorga, A. Hidalgo-Miranda, N. Gutiérrez-Nájera, and E. Hernández-Lemus, "The role of master regulators in the metabolic/transcriptional coupling in breast carcinomas," *PloS one*, vol. 7, no. 8, p. e42678, 2012. 155, 165

[274] M. J. Potthoff and E. N. Olson, "MEF2: a central regulator of diverse developmental programs," *Development*, vol. 134, no. 23, pp. 4131–4140, 2007. 155, 165

[275] E. Hernández-Lemus, K. Baca-López, and H. Tovar, "What Makes a Transcriptional Master Regulator? A Systems Biology Approach," in *Physical Biology of Proteins and Peptides*, pp. 161–174, Springer, 2015. 155

[276] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. van Nimwegen, "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences," *Bioinformatics*, vol. 28, no. 4, pp. 487–494, 2012. 156

[277] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, "Globocan 2012 v1. 0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11. 2013," *Available from: globocan. iarc. fr*, 2014. 158

[278] K. Polyak, "Heterogeneity in breast cancer," *The Journal of clinical investigation*, vol. 121, no. 10, pp. 3786–3788, 2011. 158

[279] G. de Anda-Jáuregui, R. A. Mejía-Pedroza, J. Espinal-Enríquez, and E. Hernández-Lemus, "Crosstalk events in the estrogen signaling pathway may affect tamoxifen efficacy in breast cancer molecular subtypes," *Computational biology and chemistry*, vol. 59, pp. 42–54, 2015. 158, 162

[280] J. Espinal-Enriquez, C. Fresno, G. Anda-Jáuregui, and E. Hernández-Lemus, "Rna-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer.," *Scientific reports*, vol. 7, p. 1760, May 2017. 158, 176

[281] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000. 158, 159

[282] M. Guedj, L. Marisa, A. De Reynies, B. Orsetti, R. Schiappa, F. Bibeau, G. Macgrogan, F. Lerebours, P. Finetti, M. Longy, *et al.*, "A refined molecular taxonomy of breast cancer," *Oncogene*, vol. 31, no. 9, pp. 1196–1206, 2012. 158

[283] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. Van't Veer, and C. M. Perou, "Concordance among gene-expression–based predictors for breast cancer," *New England Journal of Medicine*, vol. 355, no. 6, pp. 560–569, 2006. 158

[284] Z. Hu, C. Fan, D. S. Oh, J. S. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, *et al.*, "The molecular portraits of breast tumors are conserved across microarray platforms," *BMC genomics*, vol. 7, no. 1, p. 96, 2006. 158

[285] O. Metzger-Filho, Z. Sun, G. Viale, K. N. Price, D. Crivellari, R. D. Snyder, R. D. Gelber, M. Castiglione-Gertsch, A. S. Coates, A. Goldhirsch, *et al.*, "Patterns of recurrence and outcome according to breast cancer subtypes in lymph node–negative disease: Results from international breast cancer study group trials viii and ix," *Journal of Clinical Oncology*, pp. JCO–2012, 2013. 158, 159

[286] N. D. Arvold, A. G. Taghian, A. Niemierko, R. F. A. Raad, M. Sreedhara, P. L. Nguyen, J. R. Bellon, J. S. Wong, B. L. Smith, and J. R. Harris, "Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy," *Journal of Clinical Oncology*, vol. 29, no. 29, pp. 3885–3891, 2011. 158

[287] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10869–10874, 2001. 159

[288] R. Haque, S. A. Ahmed, G. Inzhakova, J. Shi, C. Avila, J. Polikoff, L. Bernstein, S. M. Enger, and M. F. Press, "Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades," *Cancer Epidemiology Biomarkers & Prevention*, vol. 21, no. 10, pp. 1848–1855, 2012. 159

[289] H. J. Burstein, "The distinctive nature of her2-positive breast cancers," *New England Journal of Medicine*, vol. 353, no. 16, pp. 1652–1654, 2005. 159

[290] X. R. Yang, J. Chang-Claude, E. L. Goode, F. J. Couch, H. Nevanlinna, R. L. Milne, M. Gaudet, M. K. Schmidt, A. Broeks, A. Cox, *et al.*, "Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the breast cancer association consortium studies," *Journal of the National Cancer Institute*, vol. 103, no. 3, pp. 250–263, 2011. 159, 164

[291] S. Bayraktar and S. Glück, "Molecularly targeted therapies for metastatic triple-negative breast cancer," *Breast cancer research and treatment*, vol. 138, no. 1, pp. 21–35, 2013. 159

[292] K. D. Voduc, M. C. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke, "Breast cancer subtypes and the risk of local and regional relapse," *Journal of Clinical Oncology*, vol. 28, no. 10, pp. 1684–1691, 2010. 159

[293] P. Singha, S. Pandeswara, M. Venkatachalam, and P. Saikumar, "Abstract p2-06-08: Interplay of smad2 and smad3 during tgf-β induced tmepai/pmepa1 mediated triple negative breast cancer cell growth," *Cancer Research*, vol. 76, no. 4 Supplement, pp. P2–06, 2016. 159

[294] J. Espinal-Enríquez, J. M. Siqueiros-García, R. García-Herrera, and S. A. Alcalá-Corona, "A literature-based approach to a narco-network," in *Lecture Notes in Computer Science*, pp. 97–101, Springer International Publishing, 2015. 160

[295] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A r/bioconductor package for inferring large transcriptional networks using mutual information.," *BMC bioinformatics*, vol. 9, p. 461, Oct. 2008. 160

[296] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren, J. E. L. Wong, E. T. Liu, J. Bergh, V. A. Kuznetsov, and L. D. Miller, "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer.," *Cancer research*, vol. 66, pp. 10292–10301, Nov. 2006. 160

[297] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedrén, and J. Bergh, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.," *Breast cancer research : BCR*, vol. 7, no. 6, pp. R953–64, 2005. 160

[298] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. M. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, C. Sotiriou, and TRANSBIG Consortium, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 13, pp. 3207–3214, June 2007. 160

[299] P. Farmer, H. Bonnefoi, V. Becette, M. Tubiana-Hulin, P. Fumoleau, D. Larsimont, G. Macgrogan, J. Bergh, D. Cameron, D. Goldstein, S. Duss, A.-L. Nicoulaz, C. Brisken, M. Fiche, M. Delorenzi, and R. Iggo, "Identification of molecular apocrine breast tumours by microarray analysis.," *Oncogene*, vol. 24, pp. 4660–4671, July 2005. 160

[300] A. J. Minn, G. P. Gupta, P. M. Siegel, P. D. Bos, W. Shu, D. D. Giri, A. Viale, A. B. Olshen, W. L. Gerald, and J. Massagué, "Genes that mediate breast cancer metastasis to lung.," *Nature*, vol. 436, pp. 518–524, July 2005. 160

[301] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.," *Journal of the National Cancer Institute*, vol. 98, pp. 262–272, Feb. 2006. 160

[302] A. Tripathi, C. King, A. de la Morenas, V. K. Perry, B. Burke, G. A. Antoine, E. F. Hirsch, M. Kavanah, J. Mendez, M. Stone, N. P. Gerry, M. E. Lenburg, and C. L. Rosenberg, "Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients.," *International journal of cancer. Journal international du cancer*, vol. 122, pp. 1557–1566, Apr. 2008. 160

[303] I. B. P. Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. A. Emran, N. H. Abdullah, and S. N. A. S. Hussain, "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the malaysian context," *Pathology-Research and Practice*, vol. 206, no. 4, pp. 223–228, 2010. 162

[304] R. Liu, X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, T. Hoey, G. Sherlock, J. Lewicki, K. Shedden, and M. F. Clarke, "The prognostic role of a gene signature from tumorigenic breast-cancer cells.," *The New England journal of medicine*, vol. 356, pp. 217–226, Jan. 2007. 162

[305] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh, "From the Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 102, pp. 13550–13555, Sept. 2005. 162

[306] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003. 162

[307] H. Tovar, R. García-Herrera, J. Espinal-Enríquez, and E. Hernández-Lemus, "Transcriptional master regulator analysis in breast cancer genetic networks," *Computational biology and chemistry*, vol. 59, pp. 67–77, 2015. 162

[308] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of clinical oncology*, vol. 27, no. 8, pp. 1160–1167, 2009. 162

[309] X. Wang, C. Terfve, J. C. Rose, and F. Markowetz, "Htsanalyzer: an r/bioconductor package for integrated network analysis of high-throughput screens," *Bioinformatics*, vol. 27, no. 6, pp. 879–880, 2011. 162

[310] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, "Causal analysis approaches in ingenuity pathway analysis," *Bioinformatics*, vol. 30, no. 4, pp. 523–530, 2014. 162, 182

[311] S. A. Alcalá-Corona, J. Espinal-Enríquez, G. De Anda Jáuregui, and E. H.-L. Hernández-Lemus, "The hierarchical modular structure of her2+ breast cancer network," *Frontiers in Physiology*, vol. 9, 2018. 164, 192

[312] V. Sartorelli, J. Huang, Y. Hamamori, and L. Kedes, "Molecular mechanisms of myogenic coactivation by p300: direct interaction with the activation domain of MyoD and with the MADS box of MEF2C.," *Molecular and cellular biology*, vol. 17, no. 2, pp. 1010–1026, 1997. 165

[313] M. Koch, F. Hussein, A. Woeste, C. Gründker, K. Frontzek, G. Emons, and T. Hawighorst, "Cd36-mediated activation of endothelial cell apoptosis by an n-terminal recombinant fragment of thrombospondin-2 inhibits breast cancer growth and metastasis in vivo.," *Breast cancer research and treatment*, vol. 128, pp. 337–346, July 2011. 176

[314] R. Sun, J. Wu, Y. Chen, M. Lu, S. Zhang, D. Lu, and Y. Li, "Down regulation of thrombospondin2 predicts poor prognosis in patients with gastric cancer.," *Molecular cancer*, vol. 13, p. 225, Sept. 2014. 176

[315] L. E. Kelemen, F. J. Couch, S. Ahmed, A. M. Dunning, P. D. P. Pharoah, D. F. Easton, Z. S. Fredericksen, R. A. Vierkant, V. S. Pankratz, E. L. Goode, C. G. Scott, D. N. Rider, X. Wang, J. R. Cerhan, and C. M. Vachon, "Genetic variation

in stromal proteins decorin and lumican with breast cancer: investigations in two case-control studies.," *Breast cancer research : BCR*, vol. 10, p. R98, 2008. 176

[316] C. Bonnans, J. Chou, and Z. Werb, "Remodelling the extracellular matrix in development and disease.," *Nature reviews. Molecular cell biology*, vol. 15, pp. 786–801, Dec. 2014. 178

[317] J. Espinal-Enriquez, S. Munoz-Montero, I. Imaz-Rosshandler, A. Huerta-Verde, C. Mejia, and E. Hernandez-Lemus, "Genome-wide expression analysis suggests a crucial role of dysregulation of matrix metalloproteinases pathway in undifferentiated thyroid carcinoma," *BMC Genomics*, vol. 16, Mar 2015. 178, 182

[318] C. M. Ghajar, H. Peinado, H. Mori, I. R. Matei, K. J. Evason, H. Brazier, D. Almeida, A. Koller, K. A. Hajjar, D. Y. R. Stainier, E. I. Chen, D. Lyden, and M. J. Bissell, "The perivascular niche regulates breast tumour dormancy.," *Nature cell biology*, vol. 15, pp. 807–817, July 2013. 179

[319] F. Verrecchia, M. L. Chu, and A. Mauviel, "Identification of novel tgf-beta /smad gene targets in dermal fibroblasts using a combined cdna microarray/promoter transactivation approach.," *The Journal of biological chemistry*, vol. 276, pp. 17058–17062, May 2001. 179

[320] C. A. Vaughan, S. P. Deb, S. Deb, and B. Windle, "Preferred binding of gain-of-function mutant p53 to bidirectional promoters with coordinated binding of ETS1 and GABPA to multiple binding sites," *Oncotarget*, vol. 5, no. 2, p. 417, 2014. 195

[321] Z. Odrowaz and A. D. Sharrocks, "The ETS transcription factors ELK1 and GABPA regulate different gene networks to control MCF10A breast epithelial cell migration," *PloS one*, vol. 7, no. 12, p. e49892, 2012. 195

[322] Y. Tong, D. Merino, B. Nimmervoll, K. Gupta, Y.-D. Wang, D. Finkelstein, J. Dalton, D. W. Ellison, X. Ma, J. Zhang, *et al.*, "Cross-species genomics identifies TAF12, NFYC, and RAD54L as choroid plexus carcinoma oncogenes," *Cancer cell*, vol. 27, no. 5, pp. 712–727, 2015. 195

[323] N. Lützner, J. D.-C. Arce, and F. Rösl, "Gene expression of the tumour suppressor LKB1 is mediated by Sp1, NF-Y and FOXO transcription factors," *PLoS One*, vol. 7, no. 3, p. e32590, 2012. 195

[324] A. E. Kottorou, A. G. Antonacopoulou, F.-I. D. Dimitrakopoulos, A. C. Tsamandas, C. D. Scopa, T. Petsas, and H. P. Kalofonos, "Altered expression of NFY-C and RORA in colorectal adenocarcinomas," *Acta histochemica*, vol. 114, no. 6, pp. 553–561, 2012. 195

[325] T. Zou, J. N. Rao, L. Liu, L. Xiao, H. K. Chung, Y. Li, G. Chen, M. Gorospe, and J.-Y. Wang, "JunD enhances miR-29b levels transcriptionally and posttranscriptionally to inhibit proliferation of intestinal epithelial cells," *American Journal of Physiology-Cell Physiology*, vol. 308, no. 10, pp. C813–C824, 2015. 195

[326] C. L. Davidson, L. E. Cameron, and D. N. Burshtyn, "The AP-1 transcription factor JunD activates the leukocyte immunoglobulin-like receptor 1 distal promoter," *International immunology*, vol. 26, no. 1, pp. 21–33, 2014. 195

[327] C.-C. Wang, S. S. Bajikar, L. Jamal, K. A. Atkins, and K. A. Janes, "A time- and matrix-dependent TGFBR3–JUND–KRT5 regulatory circuit in single breast epithelial cells and basal-like premalignancies," *Nature cell biology*, vol. 16, no. 4, pp. 345–356, 2014. 195

[328] C. Li, S. Li, D.-H. Kong, X. Meng, Z.-H. Zong, B.-Q. Liu, Y. Guan, Z.-X. Du, and H.-Q. Wang, "BAG3 is upregulated by c-Jun and stabilizes JunD," *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1833, no. 12, pp. 3346–3354, 2013. 195

[329] K. Eckhoff, R. Flurschütz, F. Trillsch, S. Mahner, F. Jänicke, and K. Milde-Langosch, "The prognostic significance of Jun transcription factors in ovarian cancer," *Journal of cancer research and clinical oncology*, vol. 139, no. 10, pp. 1673–1680, 2013. 195

[330] J. Thevenon, A. Bourredjem, L. Faivre, C. Cardot-Bauters, A. Calender, A. Murat, S. Giraud, P. Niccoli, M.-F. Odou, F. Borson-Chazot, *et al.*, "Higher risk of death among MEN1 patients with mutations in the JunD interacting domain: a Groupe dÔøetude des Tumeurs Endocrines (GTE) cohort study," *Human molecular genetics*, vol. 22, no. 10, pp. 1940–1948, 2013. 195

[331] H. Gazon, I. Lemasson, N. Polakowski, R. Césaire, M. Matsuoka, B. Barbeau, J.-M. Mesnard, and J.-M. Peloponese, "Human T-cell leukemia virus type 1 (HTLV-1) bZIP factor requires cellular transcription factor JunD to upregulate HTLV-1 antisense transcription from the 3 long terminal repeat," *Journal of virology*, vol. 86, no. 17, pp. 9070–9078, 2012. 195

[332] T. Nakayama, T. Higuchi, N. Oiso, A. Kawada, and O. Yoshie, "Expression and function of FRA2/JUND in cutaneous T-cell lymphomas," *Anticancer research*, vol. 32, no. 4, pp. 1367–1373, 2012. 195

[333] J. Huang, B. Gurung, B. Wan, S. Matkar, N. A. Veniaminova, K. Wan, J. L. Merchant, X. Hua, and M. Lei, "The same pocket in menin binds both MLL and JUND but has opposite effects on transcription," *Nature*, vol. 482, no. 7386, pp. 542–546, 2012. 195

[334] L. F. Zerbini, J. F. de Vasconcellos, A. Czibere, Y. Wang, J. D. Paccez, X. Gu, J.-R. Zhou, and T. A. Libermann, "JunD-mediated repression of GADD45α and γ regulates escape from cell death in prostate cancer," *Cell Cycle*, vol. 10, no. 15, pp. 2583–2591, 2011. 195

[335] M. Caffarel, G. Moreno-Bueno, C. Cerutti, J. Palacios, M. Guzman, F. Mechta-Grigoriou, and C. Sanchez, "JunD is involved in the antiproliferative effect of Δ9-tetrahydrocannabinol on human breast cancer cells," *Oncogene*, vol. 27, no. 37, pp. 5033–5044, 2008. 195

[336] F. Mehraein-Ghomi, E. Lee, D. R. Church, T. A. Thompson, H. S. Basu, and G. Wilding, "JunD mediates androgen-induced oxidative stress in androgen dependent LNCaP human prostate cancer cells," *The Prostate*, vol. 68, no. 9, pp. 924–934, 2008. 195

[337] A. Taylor and S. Halene, "The regulatory role of serum response factor pathway in neutrophil inflammatory response.," *Current opinion in hematology*, vol. 22, no. 1, pp. 67–73, 2015. 195

[338] Z. Liu, J. Zhang, Y. Gao, L. Pei, J. Zhou, L. Gu, L. Zhang, B. Zhu, N. Hattori, J. Ji, *et al.*, "Large-scale characterization of DNA methylation changes in human gastric carcinomas with and without metastasis," *Clinical Cancer Research*, vol. 20, no. 17, pp. 4598–4612, 2014. 195

[339] X.-H. Liao, N. Wang, D.-W. Zhao, D.-L. Zheng, L. Zheng, W.-J. Xing, W.-J. Ma, L.-Y. Bao, J. Dong, and T.-C. Zhang, "STAT3 protein regulates vascular smooth muscle cell phenotypic switch by interaction with myocardin," *Journal of Biological Chemistry*, vol. 290, no. 32, pp. 19641–19652, 2015. 195

[340] J. S. Bae, S. J. Noh, K. M. Kim, K. Y. Jang, M. J. Chung, D. G. Kim, and W. S. Moon, "Serum response factor induces epithelial to mesenchymal transition with resistance to sorafenib in hepatocellular carcinoma," *International journal of oncology*, vol. 44, no. 1, pp. 129–136, 2014. 195

[341] W. Li, J. Espinal-Enríquez, K. R. Simpfendorfer, and E. Hernández-Lemus, "A survey of disease connections for cd4+ t cell master genes and their directly linked genes.," *Computational biology and chemistry*, vol. 59 Pt B, pp. 78–90, Dec. 2015. 197