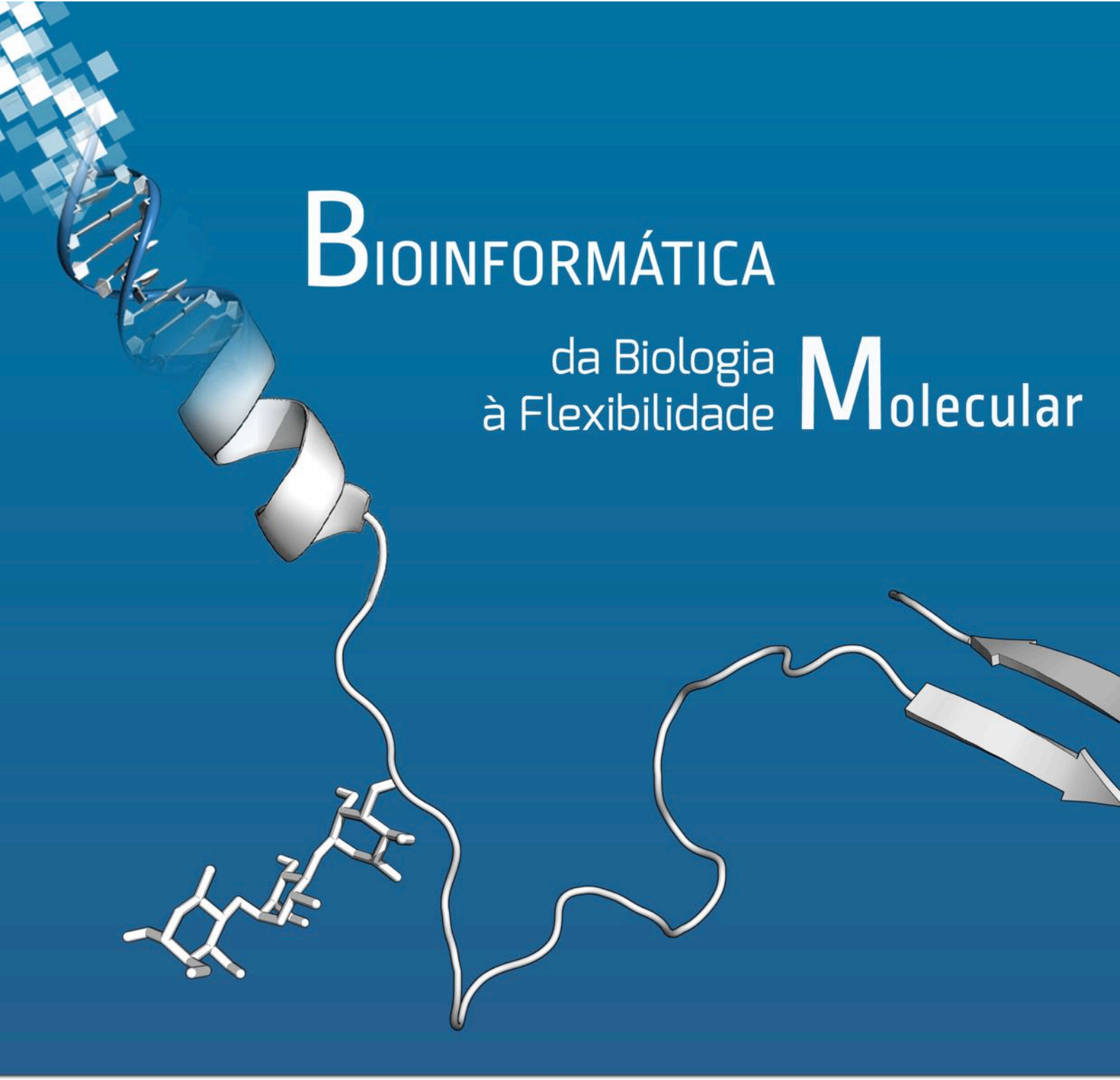


Bioinformática
da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

ISBN 978-85-69288-00-8
9 788569 288008

Sociedade Brasileira de Bioquímica e Biologia Molecular — SBBq

Apoio:
CNPq
CAPES
FAPERGS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

CDU 575.112
ISBN 978-85-69288-00-8

6. Biologia de Sistemas

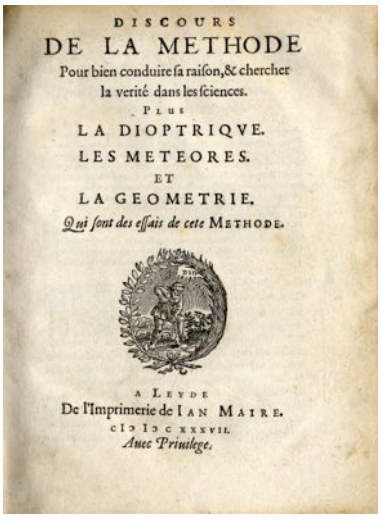
"Pensar a complexidade – esse é o maior desafio do pensamento contemporâneo, que necessita de uma reforma no nosso modo de pensar."

Edgar Morin & Jean-Louis Le Moigne

Joice de Faria Poloni
Bruno César Feltes
Fernanda Rabaioli da Silva
Diego Bonatto

6.1. Introdução

Uma das posturas metodológicas mais significativas do pensamento científico contemporâneo consiste em reduzir o todo a suas partes componentes. Por exemplo, entendemos o funcionamento de um organismo como fruto da ação de órgãos. Estes por sua vez, são compostos por tecidos, que são compostos por células. As células têm como componentes moléculas que, por fim, são compostas por átomos.


DISCOURS
DE LA METHODE
Pour bien conduire sa raison, & chercher la verité dans les sciences.
Plus
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui sont des essais de cette Methode.
A Leyde
De l'Imprimerie de Ian Maire.
cIↃ IↃ C XXXVII.
Avec Privilege.

Esta abordagem, especialmente importante e difundida na área biológica, é fruto das idéias introduzidas pelo filósofo René Descartes em meados do século XVII, indicando que cada problema encontrado deve ser dividido em tantas pequenas partes quanto for necessário para resolvê-lo de maneira mais parcimoniosa.

É neste contexto que emerge a divisão disciplinar no estudo da natureza. Desde os tempos da escola até a universidade, o conhecimento a ser ensinado manifesta-se na separação das disciplinas. Por exemplo, no meio acadêmico observamos a biologia compartimentada em botânica, zoologia, ecologia, genética, biologia celular e essas, por sua vez, subdivididas em outras áreas. Como aspecto positivo, o estudo das partes forma especialistas e divide o trabalho, facilitando o entendimento de suas partes componentes. Contudo, neste processo tem-se uma redução da complexidade característica dos fenômenos naturais, o que pode comprometer nossa capacidade de entendê-los.

De fato, a complexidade é inerente à biologia, ao funcionamento do nosso organismo e à natureza. Há a necessidade, assim, da construção de uma abordagem que inclua esta complexidade, de forma sistêmica; que interligue as diversas interações presentes e que, ao confrontá-las, consiga encontrar relações mais informativas e completas.

A partir desta premissa, emergem na década de 1950 as primeiras concepções sobre a Biologia de Sistemas (BS). Essa área, pautada nos conceitos de sistema e de complexidade, envolve um estudo sistemático de interações em um sistema biológico.

O conceito de sistema é entendido como um conjunto de partes ou elementos que possuem relações entre si, relações estas

que diferem-se daquelas realizadas com outros elementos, fora do sistema. Já a idéia de complexidade é definida como a condição de elementos de um sistema e a relação entre esses elementos em um determinado momento.

Um sistema complexo, por conseguinte, é um sistema composto de partes interconectadas que, como um todo, exibe uma ou mais propriedades que não seriam observadas a partir das propriedades dos componentes individuais, possibilitando assim a observação de novos fenômenos. Portanto, a BS é um campo que investiga as interações entre os componentes de um sistema biológico, buscando contribuir no entendimento de como estas interações influenciam a função e o comportamento do sistema.

A busca da compreensão da biologia em nível de sistema é um tema recorrente na comunidade científica. Norbert Wiener, em 1948, foi um dos proponentes da abordagem sistemática que levou ao nascimento da cibernética, ou biocibernética, consolidada com os estudos do médico neurologista, William Ross Ashby (1903-1972). A partir de 1959, Robert Rosen, sob orientação do professor Nicolas Rashevsky, propôs uma metodologia baseada na "biologia relacional", onde o mais importante na biologia era o estudo da vida em si. Após 20 anos, Ludwig von Bertalanffy (1901-1972) criou a teoria geral dos sistemas, tornando-se o precursor da BS. Em 1966 foi formalizado o estudo da BS, com o lançamento da disciplina "Teoria e Biologia de Sistemas" pelo teórico de sistemas Mihajlo Mesarovic (1928).

A partir do trabalho destes pesquisadores, a teoria geral dos sistemas pode ser definida como a área que estuda a organização abstrata de fenômenos, investigando todos os princípios comuns a todas as entidades complexas (não somente biológicas) e os modelos que podem ser utilizados para a sua descrição.

Com o avanço da biologia molecular nas décadas que se seguiram, juntamente com o nascimento da genômica funcional, grandes quantidades de dados tornaram-se disponíveis e os bancos de dados e ferramentas de análise adaptaram-se ao volume crescente de informações, permitindo construir modelos mais amplos, capazes de lidar com aspectos e fenômenos inacessíveis até então. Assim em 2000, quando o Instituto de Biologia de Sistemas foi fundado, a biologia de sistemas emergiu como um campo próprio, estimulado pelo aumento de dados "ômicos" e pelos avanços da parte experimental e da bioinformática visando o entendimento sistemático da biologia. Desde então, grupos de pesquisas dedicados à BS têm sido formados em todo o mundo.

Para tal, a BS depende de ferramentas interdisciplinares para obter, integrar e analisar diversos tipos de dados, exemplificados na Tabela 1-6. Essa abordagem requer novas técnicas de análise, ferramentas de informática, métodos experimentais e uma nova postura metodológica, articulando partes normalmente estudadas separadamente.

6.2. Biologia de Sistemas

Em suas análises, a BS relaciona partes individuais de um sistema como representações gráficas de conjuntos de nós ou vértices (V), conectados entre si por conectores ou arestas (E, do inglês *edge*). Os nós podem representar indivíduos, proteínas ou mesmo lugares, enquanto que os conectores representam a conexão que está presente entre cada par de nós. Esta representação gráfica é denominada de rede.

Muitos exemplos de rede podem ser citados, como redes de cadeia alimentar, amplamente aplicadas na ecologia, redes neurais e de interação proteica usadas na biologia e ciências médicas, além da própria *World Wide Web*, que representa uma das maiores redes funcionais no mundo da comunicação e informática.

A análise matemática de redes é denominada de teoria de grafos, e consiste em um dos principais objetos de estudo da matemática discreta. Desta forma, o termo "rede" representa as interações funcionais de um sistema, enquanto que o termo "grafo" enfa-

Tabela 1-6: Ferramentas utilizadas no estudo da BS.

Área	Tipo de análise
Bioinformática	Funções biológicas por meio de ferramentas da informática
Genômica	Sequências de DNA
Transcriptômica	Transcritos
Proteômica	Proteínas
Interatômica	Interações proteicas
Interferômica/ microRNômica	RNAi/miRNA
Epigenômica	Modificações na cromatina e no DNA
Metabolômica	Metabólitos
Fluxômica	Alterações dinâmicas de moléculas dentro de uma célula ao longo do tempo
Biômica	Bioma
Glicômica	Totalidade de carboidratos
Farmacogenômica	Genes que definem o comportamento da droga
Nutrigenômica	Relação entre a dieta e os genes individuais
Toxicogenômica	Estrutura e atividade do genoma e os efeitos biológicos adversos na exposição a xenobióticos
Imunômica	Função molecular associada aos transcritos de RNAm relacionados à resposta imune

tiza as análises matemáticas deste sistema. Neste capítulo, contudo, usaremos ambos os termos como sinônimos.

Historicamente, a teoria de grafos foi desenvolvida em 1736 pelo matemático suíço Leonard Euler na resolução do problema das sete pontes de Königsberg, atualmente conhecida como Kaliningrado, na Rússia. A cidade de Königsberg é atravessada pelo Rio Pregel e consiste de duas grandes ilhas que eram conectadas entre si e com as margens opostas por sete pontes (Figura 1A-6). O problema apresentado a Euler consistia em descobrir como caminhar pela cidade atravessando cada ponte apenas uma vez. A técnica desenvolvida pelo matemático suíço foi adaptar o mapa de Königsberg, transformando as margens e ilhas em nós e as pontes em conectores (Figura 1B-6). Euler submeteu a rede que desenvolveu a análises matemáticas, porém não encontrou solução para o problema. Contudo, a metodologia de análise de Euler foi um marco histórico na análise de problemas combinatórios, além de estabelecer o conceito de topologia que é usado em BS (ver adiante).

O emprego da teoria de grafos e suas aplicações têm apresentado um crescimento explosivo devido a sua multidisciplinaridade e ao seu conceito de modelo que permite estudar um objeto específico sem negligenciar o meio em que este objeto se encontra. Por exemplo, é possível estudar determinado fármaco considerando a atividade que diversos compostos e enzimas poderiam exercer sobre ele. Nesses estudos pode-se construir uma rede onde os nós representam compostos e enzimas e os conectores representam se há ou não relação entre eles, permitindo analisar:

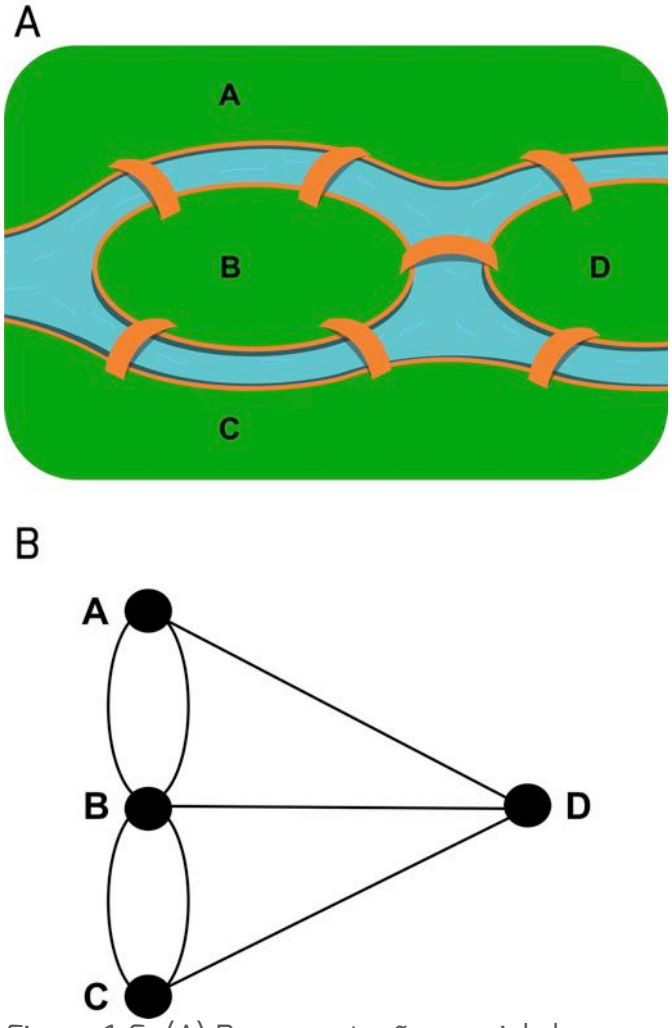


Figura 1-6: (A) Representação parcial do mapa de Königsberg e suas setes pontes. (B) Ilustração da rede desenvolvida por Euler.

i) a conectividade dos compostos ou enzimas, ou seja, que tipo de relação duas moléculas aleatórias podem apresentar na rede;
ii) a centralidade, que caracteriza as moléculas que apresentam maior influência sob a ação do fármaco em questão.

Conceitos básicos de grafos

Considerando-se a estreita relação entre a BS e a teoria de grafos, alguns conceitos matemáticos podem nos ajudar a entender e empregar esta área do conhecimento com maior domínio e propriedade. Assim, prosseguiremos com uma breve introdução sobre teoria de grafos e estrutura de rede, apresentando alguns descritores matemáticos frequentemente empregados em BS.

Uma rede (ou grafo) $G = (V, E)$ representa uma combinação de nós (V) e conectores (E) que ligam os nós. Em uma rede, o conjunto de seus nós é denotado por $V(G)$, enquanto o conjunto de seus conectores por $E(G)$. Dessa forma, o número total de nós em G é representado por n, e o número total de conectores é representado por m:

$$n(G) = |V(G)| \text{ e } m(G) = |E(G)|$$

Adicionalmente, conforme apresentado na Figura 2A-6, um conector E deve apresentar suas extremidades ligadas aos nós a e b ($a \in V$ e $b \in V$), sendo chamado eab, $E(a, b)$ ou apenas ab. Este conector pode ser representado da seguinte forma:

$$E = \{(a, b) \mid a, b \in V\}$$

As redes podem apresentar conectores diretos, ou seja, um conector orientado em determinada direção (exemplo $a \rightarrow b$, $b \rightarrow c$), sendo assim chamadas de redes direcionadas

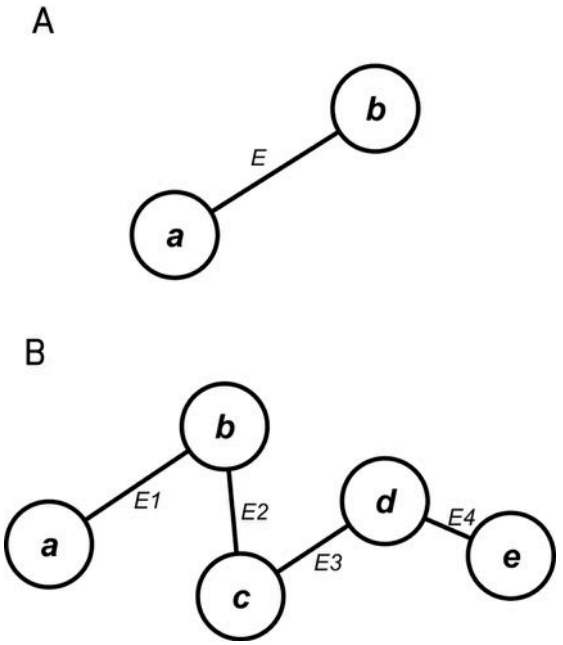


Figura 2-6: Em (A) a representação da interação de dois nós vizinhos ($V = a, b$) conectados pelo conector $E(a, b)$. Em (B) a rede pode ser descrita como $V = \{a, b, c, d, e\}$ e $E = \{ab, bc, cd, de\}$, com $n = 5$ (5 nós de a a e) e $m = 4$ (4 conectores de 1 a 4).

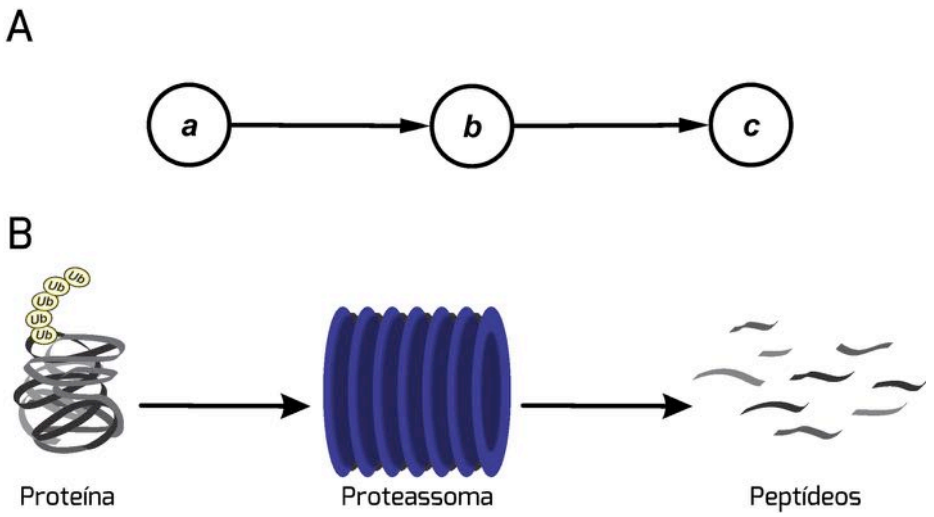


Figura 3-6: (A) Rede direta; (B) Representação da via de degradação ubiquitina-proteassoma, um dos inúmeros tipos de redes direcionadas encontradas em sistemas biológicos.

ou dígrafos (Figura 3A-6). Nos conectores E = (a, b) e E = (b, c), podemos dizer que a é antecessor a b, e b é antecessor a c. Da mesma forma, b é sucessor de a e c é sucessor de b. Um dígrafo é definido por G = (V, E, f), sendo f uma função que associa cada elemento E a um par ordenado de nós em V. Uma rede representando os mecanismos de degradação ubiquitina-proteassoma de uma determinada proteína pode ser um exemplo de rede direta após o reconhecimento da proteína ubiquitinada por proteassomas, uma vez que não é possível reverter a degradação da proteína (Figura 3B-6).

Podem também existir redes não direcionadas (Figura 4A-6), que apresentam conectores orientados em ambas as direções ($a \leftrightarrow b$, $b \leftrightarrow c$), não sendo possível assim estabelecer antecessor ou sucessor. Um exemplo típico seria a reação reversível de um substrato A para um substrato B em uma via metabólica como, por exemplo, a formação de

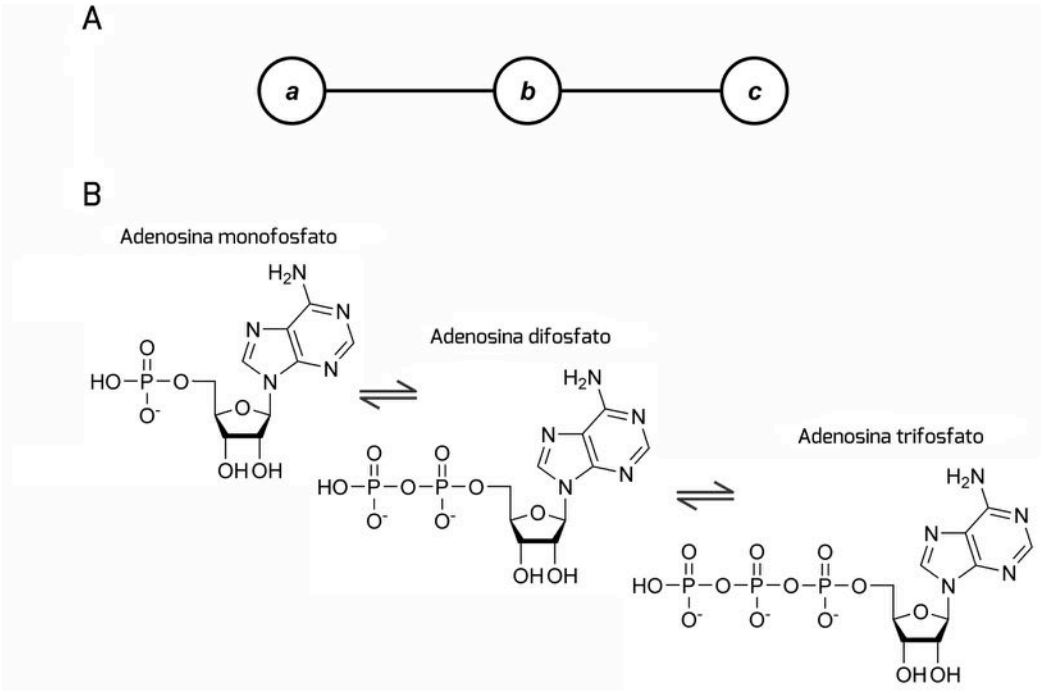


Figura 4-6: (A) Rede não direcionada; (B) Reação reversa de fosforilação e desfosforilação de adenosina difosfato, representando um exemplo de redes não direcionadas em sistemas biológicos.

diferentes moléculas fosforiladas de adenosina conforme a reação AMP↔ADP↔ATP (Figura 4B-6).

Em alguns casos, podem existir dois ou mais conectores que ligam os mesmos nós na rede. Esse tipo de interação é chamado multiconector, onde diferentes informações são representadas por cada conector, caracterizando assim um multidígrafo (Figura 5-6).

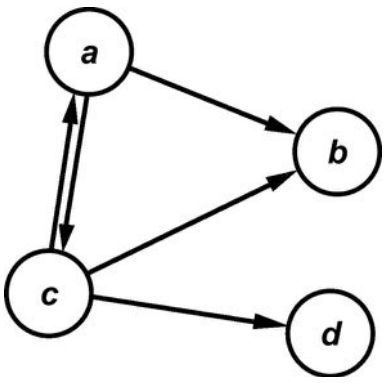


Figura 5-6: Multidígrafo $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, ac, ca, cb, cd\}$.

Observa-se, assim, que as redes apresentam interações entre os nós e que essas interações são delimitadas pelos conectores. Portanto, se $E = (a, b)$, logo os nós a e b são vizinhos ou adjacentes, e $E(a, b)$ é incidente aos nós a e b, lembrando que $E(a, b)$ se refere ao conector.

Uma das formas de representar e descrever tais interações entre os nós de uma determinada rede envolve o uso de matrizes. Assim, se considerarmos uma rede G contendo os nós v_a, ... v_n a matriz que descreve os elementos adjacentes em G é dada por:

$$a_{ij} = \begin{cases} 1 \text{ se } viv_j \in E(G) \\ 0 \text{ se } v_iv_j \notin E(G) \end{cases}$$

As tabelas representadas na Figura 6-6 são um mecanismo visual para compreender como a matriz de uma rede é elaborada, tanto para redes não direcionadas (Figura 6A-6) quanto direcionadas (Figura 6B-6).

Para as redes não direcionada (Figura 6A-6) e direcionada (Figura 6B-6), as matrizes são representadas abaixo:

$$M = \begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{matrix}$$

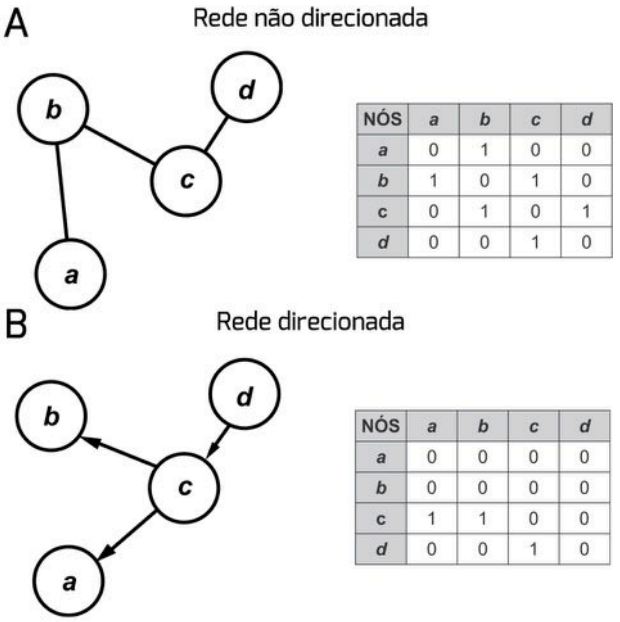
Rede direcionada

$$M = \begin{matrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix}$$

Rede não direcionada

Ao analisarmos uma matriz devemos considerar cada nó como uma coluna e uma linha distinta. Na análise da primeira matriz iremos interpor o nó representado na linha 1 (nó a) com o nó representado na coluna 1 (nó a) da mesma forma que as tabelas representadas na Figura 6-6, e como não há interação de a com a, nos referimos como 0. Da mesma forma, se consideramos a linha 1 (nó a) e a coluna 2 (nó b), há conexão, sendo representado por 1. Perceba que as matrizes são diferentes na rede direcionada e não direcionada devido à atribuição de uma conexão direcionada. Na matriz direcionada, tanto b está conectado a c quanto c está conectado a b. Contudo, na matriz não direcionada, somente c está conectado a b.

Também podemos definir uma rede como completa se $\boldsymbol{E}(G) = V(G)^{(2)}$, isto é, se dois nós selecionados aleatoriamente na rede G são adjacentes. Assim, uma rede completa tem n nós e é representada por K_n, sendo o número de conectores em K_n representado por $\binom{n}{2}$.



NÓS	a	b	c	d
a	0	1	0	0
b	1	0	1	0
c	0	1	0	1
d	0	0	1	0

NÓS	a	b	c	d
a	0	0	0	0
b	0	0	0	0
c	1	1	0	0
d	0	0	1	0

Figura 6-6: (A) Rede não direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ab, bc, cd\}$ ou $E = \{ba, cb, dc\}$, representados também na tabela pelo número 1, que indica a presença de um conector entre dois nós, exemplo $E = \{ab, ba\} = 1$. A ausência do conector entre dois nós é representada por 0. (B) Rede direcionada $G = (V, E)$, onde $V = \{a, b, c, d\}$ e $E = \{ca, cb, dc\}$. Neste caso, a tabela de interações muda devido ao direcionamento das conexões, por exemplo $E = \{ca\} = 1$, mas $E = \{ac\} = 0$.

O conjunto de nós e conectores de uma rede pode ser apresentado em uma representação mais complexa e informativa, agregando pesos (atributos) associados aos nós e conectores (Figura 7-6). Redes que apresentam nós e conectores com atributos são chamadas de redes ponderadas (G,w), onde $G = (V,E)$ e $w = V, E \in R$, sendo R o conjunto dos números reais e w correspondente à função atributo. Por exemplo, pode-se representar uma rede neural onde o atributo indica a distância que um sinal neural deve percorrer em relação ao local de origem. Assim, se P é uma trajetória na rede, $w(P)$ é considerada a extensão de P. Redes ponderadas são amplamente usadas na bioinformática, onde $G,w(a,b)$ pode representar a quantidade e a fidelidade de informações armazenadas em bancos de dados a respeito da interação entre a e b (Figura 7-6).

Também podemos nos referir a uma rede como bipartida (Figura 8-6) onde, em $G = (V, E)$, V pode ser dividido em V_x e V_y. Assim, cada nó de V_x é adjacente aos vértices de V_y. Desta forma, se consideramos E (a, b) significa que $a \in V_x$, enquanto que $b \in V_y$ ou $a \in V_y$ e $b \in V_x$. A aplicação de redes bipartidas na modelagem de redes biológicas pode ser vista em vários contextos, desde a análise de genótipos e SNPs (*single-nucleotide polymorphism*) em diferentes populações até a representação de conexões ecológicas e reações enzimáticas em vias metabólicas.

O modelo de redes visto até agora, na qual um conector se liga a dois nós, apesar de amplamente utilizado na avaliação da conectividade de redes biológicas, pode ser uma representação simplista quando se trata de redes metabólicas. A organização biológica que caracteriza as redes metabólicas em um contexto bioquímico consiste de complexas interações, frequentemente envolvendo diversos substratos e produtos. Para melhor representar a complexidade de reações bioquímicas, usam-se redes conhecidas como hipergrafos (Figura 9-6).

Os hipergrafos são caracterizados pela presença de hipervértices, que conectam mais de dois nós com propriedades distintas (Figu-

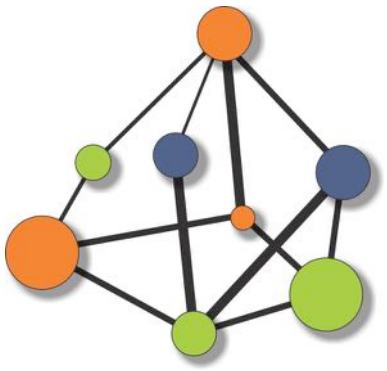


Figura 7-6: Representação de uma rede ponderada descrevendo: *i*) diferentes tipos de nós, onde cada cor representa diferentes famílias de proteínas (por exemplo, os nós verdes representam serina/treonina cinases, nós azuis representam cinases dependentes de ciclinas e nós laranjas representam as tirosina cinases); *ii*) diferentes tamanhos de nós, com atributo $w(a)$, representando o número de artigos w que citam a proteína a; e *iii*) a espessura do conector y, representando a fidelidade w da interação entre duas proteínas distintas.

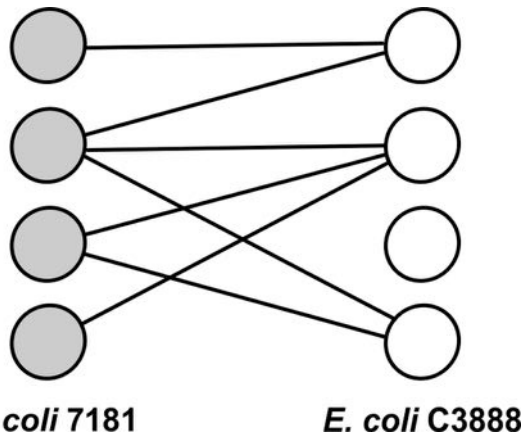


Figura 8-6: Representação de uma rede bipartida, onde os nós cinzas e brancos representam diferentes grupos de uma análise. Por exemplo, cada grupo pode representar duas linhagens diferentes de *E. coli*. Para avaliar a eficiência de transformação das linhagens, estas foram divididas em quatro amostras (representadas pelos nós) e cada amostra foi incubada com diferentes plasmídeos. Os conectores apresentam os plasmídeos que obtiveram sucesso na transformação e são comuns entre as duas linhagens.

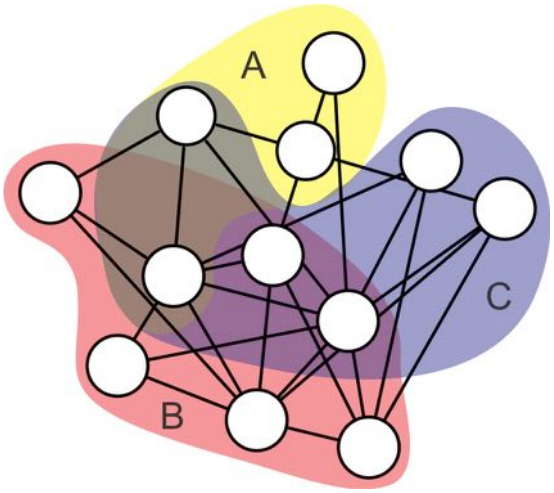


Figura 9-6: Representação de um hipergrafo. As regiões destacadas em várias cores caracterizam as diferentes propriedades ou atividades bioquímicas representadas na rede. Assim, cada cor estaria representando diferentes vias metabólicas (A, B e C). Os nós da rede indicam componentes presentes em cada uma das vias metabólicas e/ou participando de vias distintas nas regiões intersectadas.

ra 9-6). Assim, os hipergrafos são frequentemente usados em organizações bioquímicas, devido à intersecção de componentes com atividades em diferentes rotas metabólicas.

Geralmente, as redes biológicas são extensas, apresentando um grande número de nós. Contudo, análises estatísticas indicam que, dentro de uma rede maior (Figura 10A-6), podem existir redes menores que participam da composição geral e possuem maior conectividade entre si quando comparados à rede maior (Figura 10B-6). Essas subredes de $G = (V, E)$ nada mais são que uma rede $G_1 = (V_1, E_1)$, onde $V_1 \subseteq V$ e $E_1 \subseteq E$.

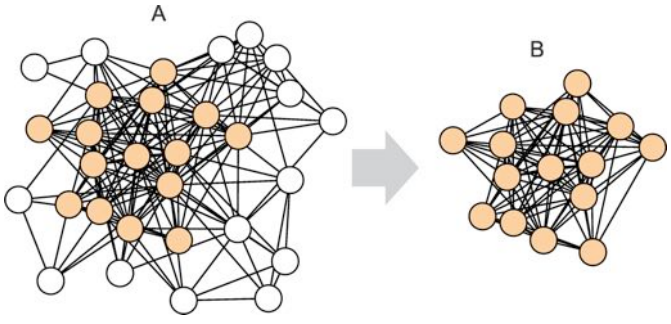


Figura 10-6: (A) Rede de interações proteína-proteína representando em laranja a subrede, o qual foi destacada em (B).

6.3. Estrutura de redes

Uma das características de uma rede é sua conectividade (também referida como grau de nó), sendo a conectividade total de uma rede definida por $C = E / N (N - 1)$, onde E representa o número de conectores e N o número total de nós.

Considere os nós V_a e V_e de uma rede. Representamos como um dos possíveis caminhos de V_a a V_e os vértices V_b, V_c e V_d, formando um conector a cada dois vértices sucessivos, caracterizados por E_1, E_8, E_4, E_7 (Figura 11-6). O nó que originou o caminho é chamado de nó inicial, enquanto que o último nó do caminho é chamado de nó final. Um caminho onde o nó inicial coincide com o nó final, sem repetições de conexões intermediárias, é chamado de circuito. Usando a mesma rede da Figura 11-6, $<d, b, c, e, d>$ formam um circuito. O comprimento de um caminho ou circuito consiste do número de conectores que pertencem ao caminho (ou circuito) ou, no caso de uma rede ponderada, pela soma dos atributos (ou pesos) dos conectores.

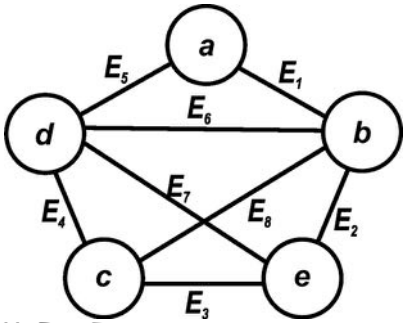


Figura 11-6: Esquema representando uma rede, onde $V = \{a, b, c, d, e\}$ e $E = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8\}$.

Um caminho de comprimento k tem exatamente $k + 1$ nós, enquanto que um circuito de comprimento k tem $k = v$ nós. Se calcularmos o comprimento de V_a a V_e, com caminho E_1, E_8, E_4, E_7 temos $k = 4$ conectores com $4 + 1$ nós. Para o circuito $<d, b, c, e, d>$ que tem como caminho E_6, E_8, E_3, E_7 temos $k = 4$ conectores, com quatro nós diferentes.

Uma importante análise em uma rede consiste em caracterizá-la conforme sua distribuição de caminhos geodésicos. Um caminho geodésico é definido como a via mais curta dentro de uma rede entre dois nós quaisquer (i e j), sendo representado por δ (i, j) em G. Um bom exemplo disso é o experimento realizado por Stanley Milgram em 1960, onde cartas foram enviadas a indivíduos aleatoriamente. A missão de cada indivíduo era enviar a sua carta a alguém que considerasse capaz de fazer com que as cartas chegassem ao seu destino final.

Essa experiência relativamente simples conclui que existem aproximadamente seis graus de separação entre dois indivíduos quaisquer no mundo. Da mesma forma, esse experimento foi a primeira demonstração significativa do efeito "mundo pequeno" (ou do inglês, *small world*), que estabelece que as redes apresentam nós conectados entre si formando um caminho mais curto entre todos os nós.

O comprimento médio de caminhos entre os nós (i, j) é definido pelo valor médio de conectores entre os nós e pode ser calculado por:

$$\delta = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{min(i,j)}$$

assumindo-se que δ_{min} (i, j) é o caminho mais curto entre os nós i e j, sendo N o número total de nós. Adicionalmente, o diâmetro da rede é definido como:

$$D = \max_{i,j} \delta_{min}(i,j)$$

e representa o maior comprimento entre dois nós. Estudos recentes têm revelado que redes biomoleculares, sociais e tecnológicas apresentam valores de comprimento médio de caminhos e diâmetro relativamente pequenos se comparados ao tamanho da rede, apresentando ordem de grandeza log (n) ou menor quando o tamanho da rede é n. Da mesma forma, a densidade de uma rede é calculada com base no número de conexões que cada nó possui, sendo definida como:

$$\rho = \frac{2m}{n(n-1)}$$

Avaliar a densidade de uma rede representa avaliar o nível de conectividade, tornando-se muito importante na definição de suas propriedades, como veremos adiante. Por exemplo, ao analisarmos a rede de interação de uma doença contagiosa, a possibilidade desta doença até então controlada tornar-se uma epidemia depende principalmente de duas variáveis: o tipo de agente infeccioso e a alta densidade de conexões (rotas de transmissão). O procedimento de quarentena (isolamento) quando um determinado indivíduo apresenta os sintomas da doença é justamente reduzir a conectividade da rede de transmissão.

Alguns modelos de rede (como as redes de livre escala e hierárquica, discutidas adiante no item 6.5.) podem apresentar clusterização, isto é, os nós tendem a se agrupar. Isso significa que se um nó A se liga ao nó B, e o nó B se liga ao nó C, então há grandes chances de A se ligar a C também. Assim, a rede é composta de centenas de triângulos, ou seja, grupos de três nós conectados entre si, onde cada lateral de um triângulo pode pertencer a outro triângulo.

Podemos quantificar a fração de triplos nós que apresentam um terceiro conector preenchendo um triângulo pelo coeficiente de clusterização:

$$C = \frac{3 \times \text{número de triângulos na rede}}{\text{número de nós triplamente conectados}}$$

Na equação, o número três presente no numerador é devido ao fato que cada lateral de um triângulo contribui com outros três triplos nós, além de garantir que C seja $0 \leq C \leq 1$. Dessa forma, o coeficiente de clusterização avalia a probabilidade dos nós i e j serem vizinhos, já que ambos são vizinhos do nó h. Assim, o coeficiente de clusterização local de um nó i pode ser determinado por:

$$C_i = \frac{2e}{k(k-1)}$$

onde um nó i tem k vizinhos com e conexões entre eles. Contudo, pode-se também atribuir o coeficiente de clusterização média para a rede total, sendo definido por:

$$C = \frac{1}{N} \sum_i C_i$$

Ao analisarmos uma rede de processos biológicos, notamos que esta apresenta um maior coeficiente de clusterização média quando comparado a uma rede aleatória. Isso possivelmente se deve ao fato de pro-

cessos celulares ocorrerem de forma dependente da organização de diversos subconjuntos (*clusters*) de biomoléculas.

Em uma rede consideramos como sendo o grau de um nó o número de conectores k que incidem a este nó. Assim, a distribuição do grau $P(k)$ é definida por ser uma fração de nós com grau k dentro de uma rede. Então sendo $k = 0, 1, 2,\ldots$ $P(k)$ indica a probabilidade de determinado nó ter grau k. A distribuição de grau é definida por:

$$P(k) = \frac{n_k}{n}$$

onde temos n nós na totalidade da rede e n_k representa a quantidade de nós com grau k.

Uma rede aleatória que apresenta n nós conectados ou não com probabilidade p, tem uma distribuição binominal de grau com parâmetros $N - 1$ e p:

$$P(k_i = k) = C_{N-1}^{k}\, p^k (1-p)^{N-1-K}$$

Outras redes, no entanto, tem distribuição de grau bem diferente. Redes de livre escala (como a maioria das redes biológicas) apresentam distribuição do grau que segue uma Lei de Potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$ (ver adiante).

Outra estimativa numérica pode ser feita, a função de distribuição cumulativa avalia a probabilidade de um nó ter um grau maior do que k:

$$P_k = \sum_{k'=k}^{\infty} p_{k'}$$

Agora, o que aconteceria se, por acaso, resolvessemos excluir alguns poucos nós da rede? Certamente iríamos alterar o comprimento de alguns caminhos e circuitos da rede de forma pouco significativa. Contudo, se formos excluindo mais nós, progressivamente, veremos que a comunicação da rede fica cada vez mais esparsa, até se tornar desconectada. A capacidade de uma rede de tolerar a deleção de nós é chamada de resiliência.

Em 2000, um estudo conduzido por Albert-László Barabási e colaboradores mostrou que a Internet pode ser altamente resiliente na remoção de nós aleatórios. Isso se deve ao fato de que a quantidade de nós com baixo grau de interação é maior em uma rede do que nós com alto grau de interação. Em compensação, se a remoção iniciar a partir dos nós com mais alto grau de interação, a alteração será brusca. Neste caso, observa-se um aumento da distância entre os nós, de forma que apenas poucos nós precisam ser removidos para destruir a comunicação da rede. Assim, fica claro que a Internet apresenta baixa resiliência na remoção de nós com alto grau, tornando-se vulnerável a ataques de *hackers*.

Outro exemplo seriam as redes de interação proteína-proteína. Estas redes geralmente apresentam muitas proteínas com poucas interações e algumas proteínas possuindo muitas interações (chamadas de *hubs*, ver adiante). Desta forma, redes de interação proteína-proteína são resilientes à deleção de nós aleatórios, porém extremamente vulneráveis a ataques em proteínas *hubs*.

Os nós de uma determinada rede podem apresentar tendências de conexão. Em outras palavras, duas redes completamente diferentes topologicamente podem apresentar a mesma distribuição do grau. Assim, em uma rede é preciso considerar o padrão de correlação do grau dos nós, onde a conectividade de um nó reflete nas suas possibilidades de ligação.

A tendência de conexão que uma rede apresenta pode ser chamada de assortatividade e desassortatividade. A assortatividade significa que os nós de uma rede apresentam uma tendência a interagirem com outros nós semelhantes, por exemplo, nós do tipo A interagem preferencialmente com nós também do tipo A (Figura 12A-6). Vértices com alto grau tendem a interagir com vértices que também apresentam alto grau. No entanto, chamamos de desassortatividade se os nós de uma rede interagem preferencialmente com nós diferentes dele mesmo, por exemplo, nós do tipo A tendem a interagir com nós do tipo B. Neste caso, um nó com alto grau tem tendência a interagir com nós que apresentem baixo grau (Figura 12B-6).

A correlação de grau dos nós i e j é feita por distribuição de probabilidade conjunta $P(k_i, k_j) = P(k_i)\, P(k_j)$. Podemos ainda calcular a assortatividade ou desassortatividade da rede como um todo, considerando:

$$r = \frac{\sum_i eii \; - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

Se $r = 1$ a rede é considerada assortativa, enquanto que se $r = -1$, a rede é completamente desassortativa.

Caracteristicamente, redes assortativas são mais resilientes e apresentam *hubs* bem conectados, enquanto que redes desassortativas são redes mais vulneráveis com nós conexos a *hubs* esparsos (Figura 12-6).

A conectividade de uma rede também pode ser avaliada pela teoria da percolação. Essa teoria tem por objetivo estudar a conectividade da rede pela avaliação de sua arquitetura, caracterizando a distribuição do tamanho dos *clusters* e descrevendo como ocorre a transferência de informações, por exemplo, de A para B.

Redes aleatórias caracteristicamente apresentam baixa tendência em possuir pequenos *clusters* isolados e uma grande probabilidade em formar um componente conectado gigante. Como visto anteriormente, determinadas redes são altamente resilentes à deleção aleatória de nós. A variação na fração dos nós no maior componente da rede (componente gigante) é a forma mais fácil de calcular a resiliência. Imagine dois nós conectados na rede. Se estes nós pertencem a um componente gigante, há grande probabilidade de se comunicarem com uma extensa proporção de nós da rede. No entanto, nós que participam de pequenos componentes comunicam-se apenas com uma parte reduzida da rede. Essa capacidade de comunicação é responsável pela forma como a informação é transferida de um ponto a outro. Assim, associamos a resiliência com a percolação local (refere-se aos nós), enquanto que a percolação de ligação (refere-se aos conectores) está relacionada ao processo de dispersão (Figura 13A-6).

Também podemos considerar os nós de uma rede como ocupados (funcionais) ou desocupados (falhos), dependendo da sua funcionalidade. A probabilidade de um nó estar ou não ocupado pode ser uniforme ou pode depender do grau do nó, sendo que os nós funcionais da rede formam o componente gigante em um modelo de percolação. Assim, os nós ou conectores falhos não participam da transferência de informação, e igualmente, não participam do componente gigante (Figura 13B-6). Dessa forma, ao observar a propri-

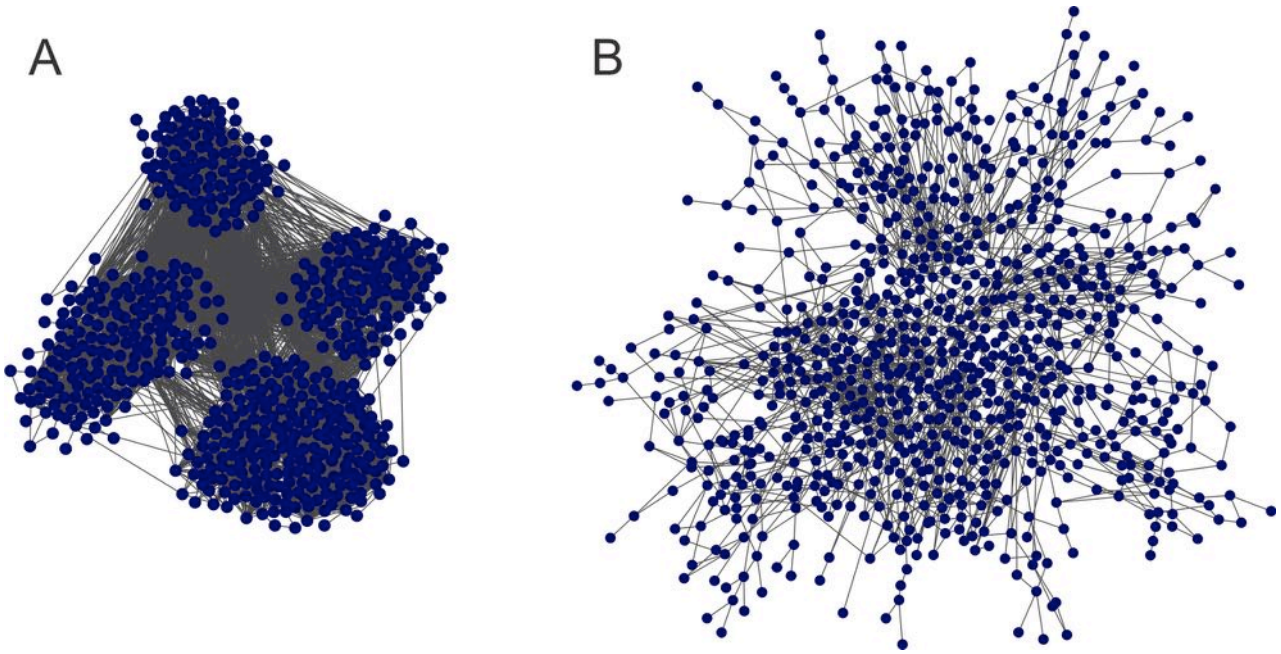


Figura 12-6: Ilustração representando em (A) uma rede assortativa com nós bem conectados que apresentam conexões com outros nós também fortemente conectados. Em (B), uma rede desassortativa, onde os poucos nós que apresentam mais conexões interagem com nós menos conectados, resultando em uma rede menos densa.

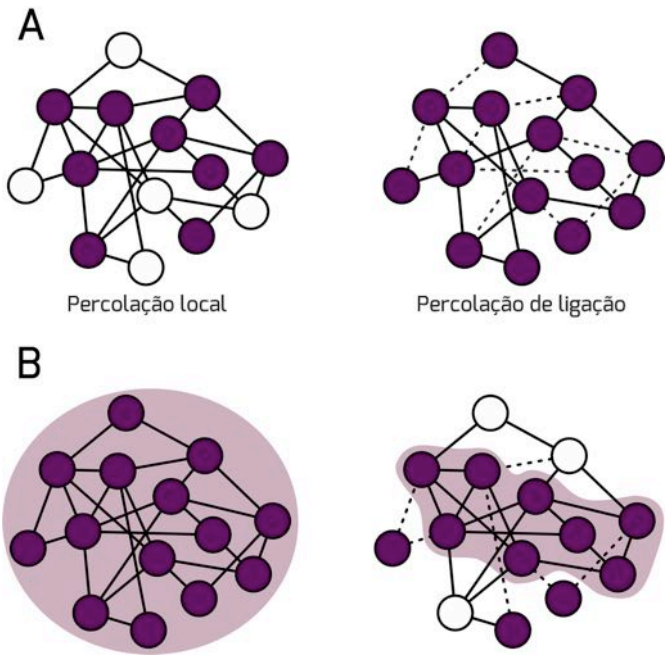


Figura 13-6: (A) Redes de percolação local e de ligação, onde os nós sólidos estão ocupados ou funcionais, enquanto que os nós brancos são desocupados ou falhos. (B) Representação do componente gigante. Após o surgimento de nós e conectores falhos, sua proporção é alterada e, por conseguinte, as possibilidades de transferência de informações.

edade de percolação de um *cluster*, considerando uma probabilidade de ocupação variável, podemos determinar que isso afeta diretamente a conectividade de uma rede, tornando-a altamente resiliente ou não. Porém, ao combinarmos a percolação local e de ligação, teremos um modelo robusto contra falhas de nós ou conectores.

Os modelos de percolação são utilizados em muitas redes, porém um dos modelos mais interessante é o da dispersão de uma doença. Nesse modelo, cada nó representa o hospedeiro e os conectores representam a capacidade de transmissão da doença entre um hospedeiro e outro. O nó (indivíduo hospedeiro) está ocupado se for suscetível à doença, enquanto que um nó que representa um indivíduo que tomou a vacina seria considerado como desocupado. Da mesma forma, os conectores são considerados ocupados se há possibilidade de transmissão (Figura 14-6). Levando em conta este modelo, o início de uma epidemia representa a transição de percolação.

Apesar de ter sido originalmente desenvolvida com o objetivo de responder às perguntas em química orgânica, os modelos de percolação têm sido usados com sucesso para estudar diversos fenômenos, como transferência de sinal em neurônios e condutividade elétrica. Em 1987, Robert H. Gardner foi um dos primeiros pesquisadores a usar a teoria de percolação na Ecologia da Paisagem, sendo útil também na avaliação de corredores ecológicos e redes de incêndios florestais.

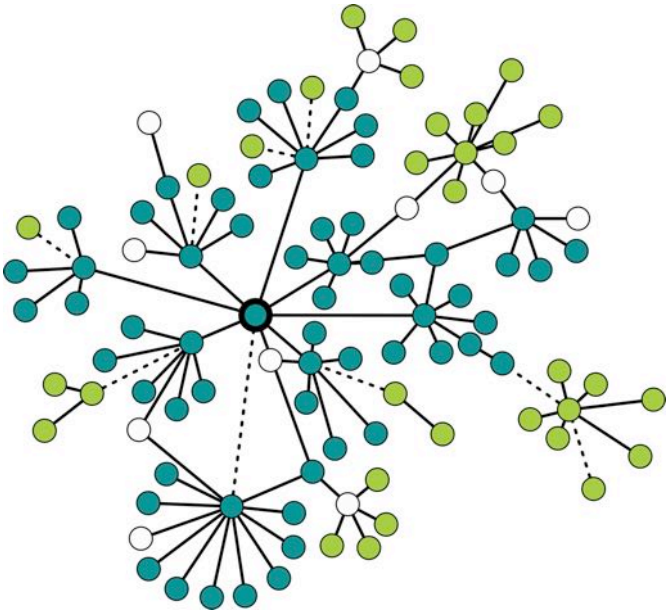


Figura 14-6: Modelo simplificado de dispersão de uma doença considerando um grupo de trabalho em uma empresa. Suponhamos que o indivíduo central contraiu uma doença viral de fácil transmissão, como a gripe simples. Assim, todos os indivíduos com os quais ele entrou em contato neste período também contraíram a doença (nós azuis), com exceção daqueles que foram vacinados (nós brancos). Neste caso, além de não contraírem a doença, também não a dispersaram. Os conectores pontilhados indicam que não houve interação física durante o período passível de contrair a doença entre o indivíduo saudável com o contaminado. Desta maneira, os indivíduos representados pelo nó verde claro, apesar de não terem sido vacinados, não contraíram a doença por não entrarem em contato com indivíduos contaminados.

6.4. Propriedades de rede

Diversas propriedades são regularmente empregadas na análise de redes biológicas, cada uma fornecendo informação sobre as interações e/ou componentes de um determinado sistema. Estas propriedades podem ser referentes a nós individuais, isto é, grau de nó ou *node degree*, ou podem contemplar a rede como um todo como é, por exemplo, o caso da modularização e do diâmetro da rede.

Em uma análise de biologia de sistemas, a análise estatística destas propriedades possui papel crítico na geração de dados conclusivos e confiáveis, constituindo-se assim em redes capazes de descrever com alto grau de fidelidade um determinado modelo biológico, de identificar alvos proteicos críticos na rede ou no desenvolvimento de caminhos moleculares.

Modularidade

Uma das principais características quando nos referimos a propriedades da topologia de redes é a chamada modularidade ou clusterização. O conceito de modularidade é antigo e já amplamente usado em outras áreas do conhecimento, como nas ciências sociais. Dentro das ciências biológicas, é um conceito comum nas áreas da biologia evolutiva, biologia molecular, biologia de sistemas e biologia do desenvolvimento.

Todas as ideias de modularidade giram em torno do conceito de padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados. De forma geral, a modularidade é um princípio de união entre diferentes tipos de elementos e conexões naturalmente formadas no meio biológico, como na interação entre indivíduos de mesma espécie. Um exemplo é a *Pollenia rudis*, uma espécie de mosca conhecida como *cluster fly* em decorrência de seu hábito de se agrupar com indivíduos da mesma espécie.

Este princípio é visto em todos os lugares, seja na nossa tendência de formar sociedades e grupos preferenciais de interação interpessoais ou na nossa tendência de organizar objetos por seu tipo, função e cores, dentre outros. Em nível molecular é visto, por exemplo, em elementos que atuam num mesmo processo biológico, como conjuntos de moléculas de RNA responsáveis pela degradação e síntese de ácidos nucleicos ou grupos de proteínas que atuam num mesmo processo biológico como a replicação de DNA e a transcrição gênica.

Existem dois tipos distintos de módulos:

i) Módulo Variacional: apresenta características que variam entre seus componentes e são relativamente independentes de outros módulos, porém possuem um número considerável de ligações com outros módulos;

ii) Módulo Funcional: possui elementos que normalmente atuam juntos em alguma função fisiológica distinta e são semiautônomos (*quasi-autonomous*) de outros módulos. Esses módulos compreendem a maioria dos módulos vistos em redes biológicas.

Módulos variacionais podem ser exemplificados na Figura 15B-6 e C, representando a formação de uma mandíbula de rato. Apesar de se tratar da diferenciação de um tecido, podemos usá-la como modelo variacional devido ao fato de diferentes proteínas e genes serem responsáveis pela formação de uma unidade estrutural única (o ramo ascendente e da região alveolar). Desta maneira, é uma unidade estrutural (um único osso) que se origina de diferentes módulos. Assim, o módulo variacional consiste numa integração de vários de genes que dividem efeitos pleiotrópicos entre si e que possuem poucos efeitos pleiotrópicos com outros *clusters*, sendo praticamente independente.

Módulos de genes de desenvolvimento embrionário, relacionados à diferenciação ou formação de padrões corporais, tendem a ser quase independentes de outros módulos, uma vez que erros na sua expressão ou atuação podem ser letais para o embrião. Por isso, esses módulos de desenvolvimento tendem a depender de elementos dentro do próprio

grupo para sua expressão. Podemos visualizar um exemplo de um módulo funcional na Figura 15A-6.

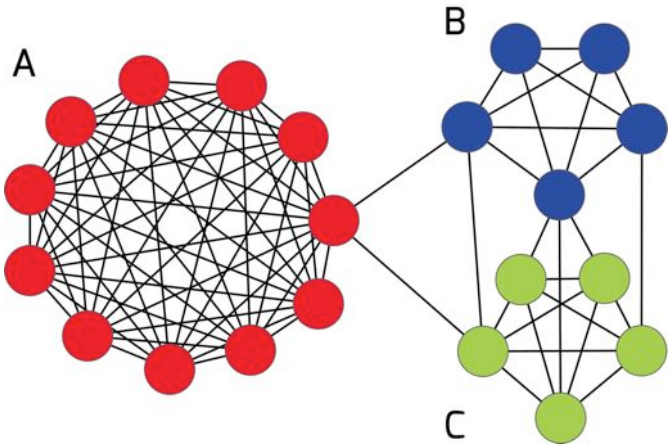


Figura 15-6: Exemplos de uma rede com diferentes módulos representados. Os módulos variacionais B (azul) e C (verde) se encontram praticamente independentes do módulo A (vermelho), porém possuem proteínas em comuns entre si. Contudo, o módulo A pode ser considerado funcional, uma vez que possui apenas uma conexão com cada outro módulo, sendo praticamente independente.

Ao determinarmos a quantidade e o tipo de módulos presentes em uma rede devemos levar em consideração o coeficiente de agrupamento (C_i) ou clusterização. O coeficiente analisa a tendência de um nó de se associar com seus vizinhos ("*cliquishness*"), onde "clique" é definido como um grafo maximamente conectado.

Como mencionado anteriormente, a clusterização é dada pela fórmula $C_i=2n/k_i(k_i-1)$, onde k_i é o tamanho da vizinhança de vértices (nós) do vértice i, e n é o número de conectores na vizinhança. Assim, quanto maior o coeficiente de clusterização, mais conectado é o *cluster*. Evolutivamente, as proteínas que compõem módulos altamente agrupados tendem a ser conservadas ou perdidas juntamente, caso haja uma variação dentro do grupo.

Outro conceito essencial para entender a formação de um *cluster* em um sistema biológico é a presença de *hubs*. Os *hubs* podem ser classificados em dois grupos:

i) *party hubs*, proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligadas no mesmo tempo e/ou espaço,

ii) *date hubs*, que são *hubs* que se ligam a diferentes proteínas em diferentes módulos (inter-módulo), ou seja, diferentes tempo e/ou espaços, consequentemente apresentando um papel global na rede (Figura 16-6). Estes termos podem ainda receber denominações específicas no contexto do conceito de centralidades (ver adiante).

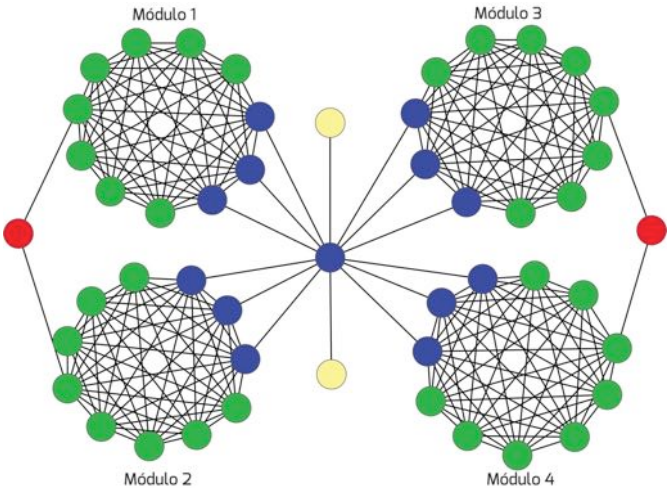


Figura 16-6: Diferentes tipos de centralidade em uma rede biológica. Em verde são apresentadas proteínas envolvidas em *party hubs* e encontradas em módulos. Em amarelo encontram-se as proteínas não-*hub*/não-gargalo, que são aquelas que não possuem alto valor de grau de nó ou *betweenness*, sendo consideradas componentes funcionais dos módulos. Em azul estão as proteínas *hub*-gargalo (*date-hub*) que possuem alto valor de grau de nó e de *betweenness*, sendo consideradas fundamentais para o funcionamento de redes. Em vermelho estão identificadas as proteínas do tipo gargalo, com alto valor de *betweenness* e essenciais na ligação entre módulos e processos biológicos.

Os *party hubs* são componentes clássicos de módulos funcionais, uma vez que estes são quase independentes de outros módulos, enquanto *date hubs* são fundamentais para módulos variacionais, pois estes se ligam a

outros módulos.

Assim, uma mutação em um *party hub* vai afetar principalmente as proteínas referentes ao seu próprio módulo, enquanto a mutação em um *date hub* (Figura 16-6) pode afetar vários módulos. Contudo, não existe diferença de importância entre *party* ou *date hub*. A deleção de um *hub* em um módulo funcional pode ser tão letal quanto a deleção em um módulo variacional.

Baseado em dados estruturais, os *hubs* podem ser ainda classificados em *singlish* (com uma ou duas interfaces) e multi-interface (com mais de duas interfaces). *Hubs* com interface *singlish* somente se ligam a outras proteínas de maneira alternada e transitória, enquanto *hubs* multi-interface se ligam a diferentes proteínas concomitantemente.

Ontologias Gênicas

Nos últimos anos, o desenvolvimento e uso de técnicas de análise como microarranjos, ChIP-chip e espectrometria de massas e suas aplicações no estudo de cada vez mais organismos gerou um grande acúmulo de dados genômicos e proteômicos. A leitura e interpretação simples e concisa destes vem requerendo o desenvolvimento de novas abordagens, contexto no qual, em 1990, foi criado o chamado *Gene Ontology Project*.

Ontologia gênica refere-se ao produto de um determinado gene e à função que ele desempenha na maquinaria celular. São classificadas em três níveis hierárquicos:

i) Componente celular, descrevendo a localização da proteína na célula;

ii) Processo biológico, referindo-se à série de eventos realizados por uma ou mais funções celulares;

iii) Função molecular, descrevendo a atividade que uma dada proteína desempenha no meio celular.

Essas informações são guardadas em forma de "anotações ontológicas", onde cada uma possui um número de identificação e se encontram disponíveis em bancos de dados como www.geneontology.org.

Da mesma forma, essas anotações não são restritas a humanos, mas abrangem diversos organismos modelo como *Mus musculus*, *Gallus gallus*, *Saccharomyces cerevisiae*, *Caenohabditis elegans* e *Escherichia coli*, além de outros organismos não-modelo mas que já possuem alguma anotação.

De um modo geral, a ontologia gênica tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um mesmo processo biológico. Em biologia de sistemas o emprego de ontologias gênicas pode se mostrar muito útil para direcionar a análise da rede, possibilitando a verificação dos tipos de processos biológicos existentes na rede e das proteínas presentes. Um modelo hipotético de como uma rede poderia se apresentar em termos de ontologias gênicas se encontra na Figura 17-6, onde diferentes nós poderiam estar relacionados a diversos processos.

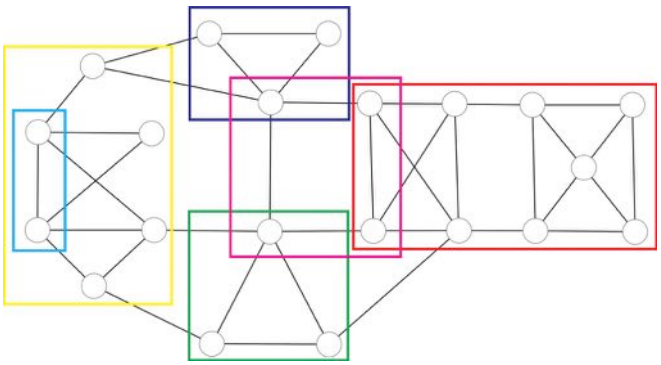


Figura 17-6: Modelo hipotético da presença de ontologias gênicas em uma rede. Na figura acima, cada cor representa um processo identificado. É importante ressaltar que uma proteína pode estar presente em mais de uma ontologia. Da mesma forma, uma ontologia pode estar dentro de outra. Como por exemplo, o quadrado amarelo poderia significar transcrição, enquanto o quadrado azul claro (inserido no amarelo) poderia significar apenas o complexo de iniciação da RNA polimerase II.

A Figura 18-6 mostra um exemplo de aplicação de ontologias gênicas em uma rede biológica. Nessa análise foi utilizado o programa *Biological Network Gene Ontology*

(BiNGO) 2.44, um *plug-in* do programa Cytoscape. É possível, assim, identificar proteínas ou genes com efeitos pleiotrópicos, a saber: a proteína Tp53, a proteína *breast cancer 1* (BRCA1) e a proteína *bloom syndrome protein* (BLM), as quais se encontram nas três ontologias da rede (reparo de DNA, regulação positiva da transcrição e ciclo celular).

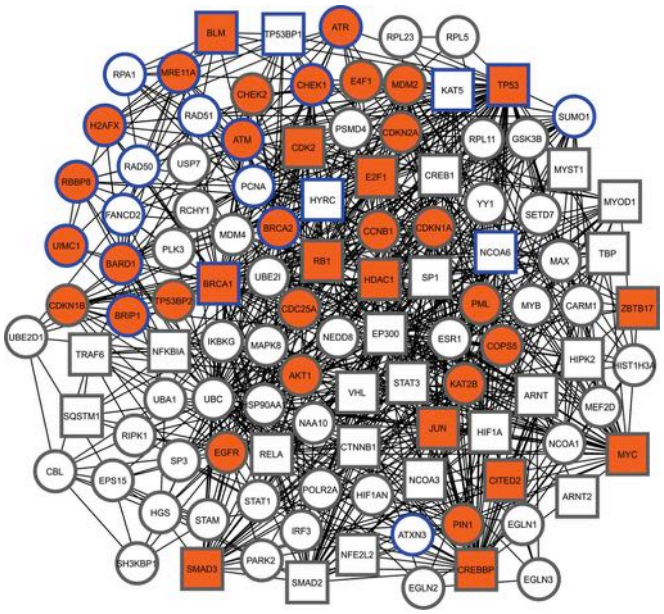


Figura 18-6: Exemplo de uma rede analisada pelo *plugin* BiNGO 2.44, o qual analisa as principais ontologias gênicas. A rede mostra três processos biológicos (GOs): *i*) Regulação do ciclo celular (nós de cor laranja); *ii*) Regulação positiva da transcrição (nós de formato quadrado); *iii*) Resposta a dano de DNA (nós com a linha azul). É possível observar que mais de um nó compõe diferentes GOs.

Centralidades para nós

Como vimos até então, a grande vantagem da biologia de sistemas é permitir a visualização dos componentes moleculares de um sistema biológico de forma dinâmica e global. Contudo, quando falamos de uma rede, temos que levar em consideração todas suas estruturas, como *hubs* e módulos. Deste modo, o objetivo da análise de centralidades é procurar o elementos mais importantes na topologia geral da rede.

Grau de nó

Um dos parâmetros básicos de análise topológica é o parâmetro de grau de nó (ou *node degree*), referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó. Esses nós que apresentam uma grande quantidade de conexões são chamados de *hubs*, os quais são conectados a outros *hubs* ou nós com menos conexões (Figura 16-6). Como veremos posteriormente, uma rede de livre escala é definida por uma lei de potenciação, o que significa que essa rede terá poucos nós altamente conectados. O grau de nó é referente ao valor distribuição de nó, $P(k)$, que informa a probabilidade de um nó ter k conexões, conforme visto em *Estrutura de redes.*

Numa visão biológica, podemos exemplificar um *hub* como uma proteína que se liga a várias outras e acaba possuindo uma função regulatória importante na rede. Normalmente, proteínas consideradas apenas *hubs* se encontram dentro de módulos. A perda de conexões de uma proteína *hub* pode lhe tirar esta condição modular. Sua deleção em uma rede de interação proteína-proteína poderia afetar a ação de diversas proteínas vizinhas e até mesmo na formação de módulos.

Betweenness

O parâmetro denominado *betweenness* é definido como o número de caminhos mais curtos que passam por um único nó, estimando a relação entre eles. Por exemplo, para calcular o valor de *betweenness* de um nó n é calculado o número de caminhos mais curtos entre i e j, e a fração deste caminhos que passam pelo nó n. Deste modo, um nó n pode ser atravessado por diversos caminhos alternativos, que ligam i e j.

Matematicamente, o valor de *betweenness* é dado pela seguinte fórmula:

$$Bet(n) = \sum_{i \neq n \neq j \in V} \frac{\sigma_{ij}(n)}{\sigma_{ij}}$$

onde σ_{ij} representam caminhos geodésicos entre os nós i e j, e $\sigma_{ij}(n)$ é o total destes caminhos mais curtos

que passam por n.

Por exemplo, uma proteína com alto valor de *betweenness* apresentaria uma elevada capacidade de interação e/ou sinalização com outras proteínas, processos biológicos ou *clusters*. Uma proteína com tais características é chamada de *bottleneck* ou gargalo. Na Figura 16-6, temos dois exemplos de uma proteína com alto valor de *betweenness*.

Não existe uma maneira óbvia de se encontrar proteínas gargalo. Porém, é possível que rotas de sinalização possuam grande incidência de proteínas gargalo, uma vez que são necessárias para sinalização entre compartimentos e processos biológicos distintos. Contudo, proteínas gargalo não necessariamente possuem um grande número de interações com outras proteínas.

Closeness

O valor de *closeness* pode ser entendido como o caminho mais curto entre um nó n e todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó (Figura 19-6). Um alto valor de *closeness* indica que todos os outros nós estão próximos do nó n, enquanto que um baixo valor indicaria que os outros nós encontram-se distantes.

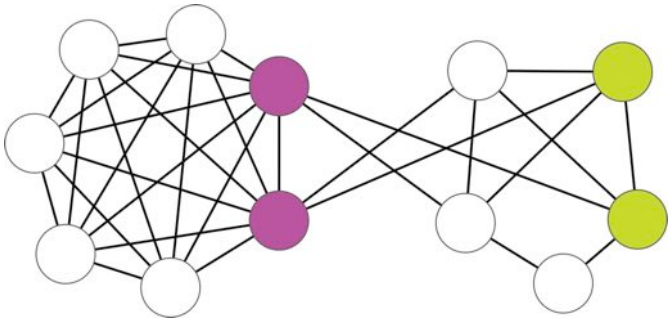


Figura 19-6: Caracterização de nós com diferentes valores hipotéticos de *closeness*. Os nós em roxo, dadas as suas maiores conectividades com a rede no geral, possuem um valor maior de *closeness*, enquanto que os nós em verde, por possuírem poucas conexões com a rede, apresentam baixo valor de *closeness*.

Este parâmetro é dado pela fórmula:

$$Clo(v) = \frac{1}{\sum w \in v^{dist\,(v'w)}}$$

onde o valor de *closeness* de um nó v [$Clo(v)$] é determinado através do cálculo e somatório dos caminhos mais curtos entre um nó v e todos outros nós w [$dist(v,w)$] dentro da rede.

Uma proteína com alto valor de *closeness* poderia ser considerada relevante para muitas proteínas, porém irrelevante para outras. Em termos biológicos, ela seria importante na regulação de muitas proteínas, porém sua atividade pode não influenciar outras. Ao compararmos essas informações com módulos podemos dizer que uma rede com uma média de *closeness* alta é mais provável de estar organizada como um módulo funcional, enquanto uma com baixo valor de *closeness* é mais provável de estar organizada como um módulo variacional.

Diâmetro

O diâmetro pode ser considerado um dos primeiros parâmetros referentes à "compactação", isto é, proximidade dos nós da rede. Ele indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo. Deste modo, uma rede com baixo diâmetro é considerada mais completa, uma vez que suas proteínas estão mais interligadas entre si.

Um baixo diâmetro pode indicar que as proteínas de uma determinada rede possuem uma maior facilidade de se comunicar e/ou influenciar umas as outras, apontando para uma relação funcional co-evolutiva (Figura 20-6).

Os parâmetros de centralidades podem ser alterados com a adição ou deleção de nós ou conexões na rede (Figura 21-6). Como já mencionado, em um sistema molecular, a perda de uma conexão pode ser considerada a mudança de um domínio, impedindo a ligação

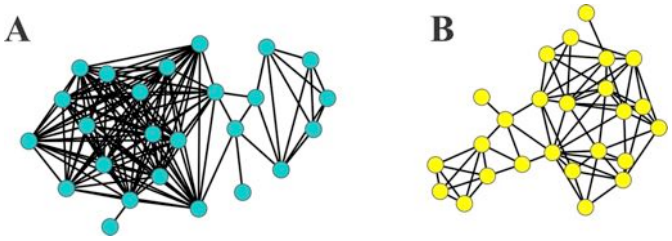


Figura 20-6: Em (A) uma rede com alto diâmetro e em (B) rede com baixo diâmetro. Pelo fato dos nós da figura A estarem mais interligados entre si, a rede é considerada mais "compacta", pois seus nós mais facilmente podem influenciar uns aos outros. Entretanto, em B, a rede possui muito menos conexões, portanto a deleção de um nó irá afetar a rede de um modo mais sutil.

de duas proteínas ou a mudança de um produto gênico, criando proteínas anormais que não mais farão as mesmas conexões. Contudo, mudanças topológicas nas redes biológicas são processos normais durante a evolução. A deleção e a duplicação de um gene, assim como a perda de interações, sejam pela mudança estrutural ou de função, são processos muitas vezes selecionados e necessários para sobrevivência celular.

Centralidade para conectores

Os elementos mais informativos de uma rede de interação podem ser avaliados através da análise da centralidade. Dentre as possíveis centralidades avaliadas, o *betweenness* de um conector pode medir a influência de certos conectores no fluxo de informações entre os componentes da rede.

O *betweenness* de um conector e é simplesmente o número de caminhos mais curtos entre pares de nós que percorrem e. Se uma rede contém módulos que são conectados por poucos conectores intermodulares, então os caminhos mais curtos entre os diferentes módulos devem passar por estes poucos conectores. Assim, os conectores unindo módulos terão altos valores de *edgebetweenness* (Figura 22-6).

Neste caso, os pares de nós unidos pelos conectores serão de diferentes módulos. Se o valor de *edgebetweenness* de um conector é baixo, esse conector provavelmente fará parte do módulo, uma vez que dentro do módulo os nós são mais interligados entre si. Portanto, *edgebetweenness* é a frequência de um conector que se coloca sobre os caminhos mais curtos entre todos os pares de nós. Em

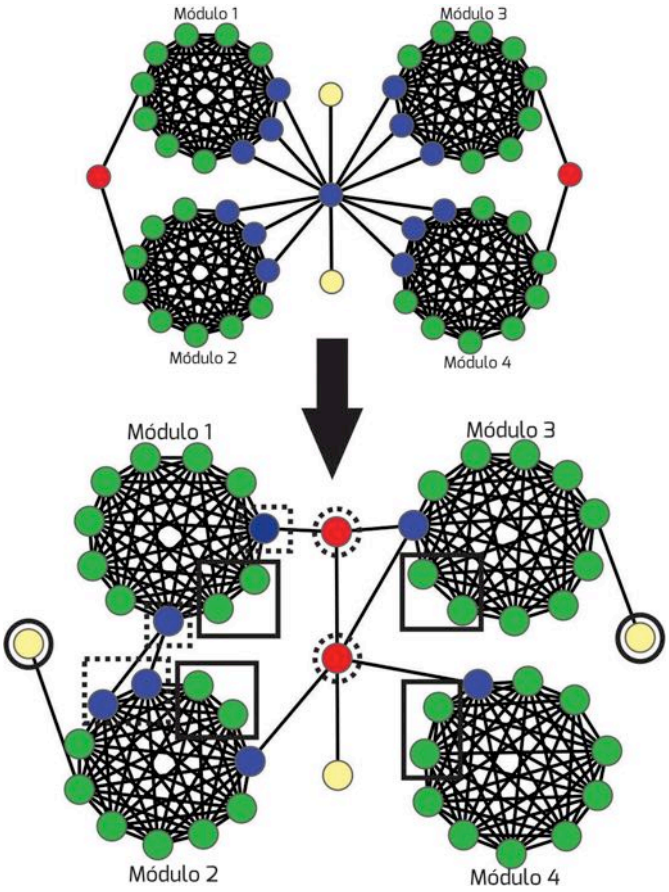


Figura 21-6: Modificações na topologia de rede podem alterar as centralidades. Devido à perda de conexões com nós fora do módulo, os nós marcados pelos quadrados foram transformados em *party-hubs* (nós verdes), deixando de ser *hubs*-gargalos (nós azuis). Porém, marcados pelos quadrados pontilhados, há nós que além de ganharem conexões, passaram a se ligar a outros módulos, saindo do estado de não-*hub*/não-gargalo para *hub*-gargalo (nós amarelos). Marcados por círculos, os nós antes gargalos (nós vermelhos), agora pela perda de uma conexão, se tornam não-*hubs*/não-gargalos. Por fim, os nós marcados pelos círculos pontilhados, devido à perda de muitas conexões (nó central) e ao ganho de uma conexão (nó acima), se tornam gargalos, perdendo os status de *hub*-gargalo e de não-*hub*/não-gargalo respectivamente.

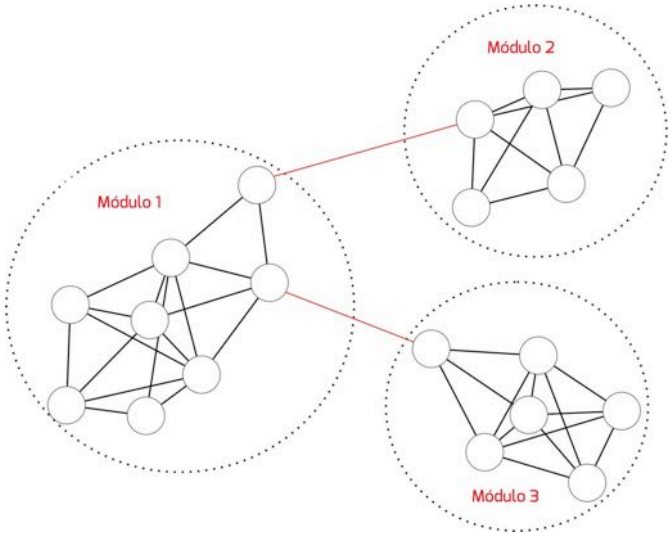


Figura 22-6: Representação de *edgebetweenness*. Conectores em vermelho apresentam valores altos de *betweenness*, pois representam o caminho mais curto do fluxo de informação entre os três módulos representados.

uma rede proteica, um conector com alto valor de *betweenness* provavelmente representa o caminho mais curto de comunicação entre dois processos biológicos.

Como conectores com altos valores de *betweenness* são mais prováveis por posicionarem-se entre módulos, a remoção sucessiva destes conectores pode eventualmente isolar estes mesmos módulos. Essa desordem na rede, conforme será visto adiante, é conhecida como perturbação de conector.

6.5. Tipos de redes

Rede Aleatória

Os matemáticos Paul Erdös e Alfréd Rényi iniciaram seus estudos sobre redes aleatórias em 1960. Este modelo de rede tem impulsionado o interesse de diversos cientistas ao longo dos anos por ser um dos primeiros modelos de rede descoberto. Porém, apesar de amplamente estudadas, redes aleatórias não capturam a realidade de um sistema biológico (Figura 23-6).

Essas redes consistem de N nós, com cada par de nós conectados (ou não) com

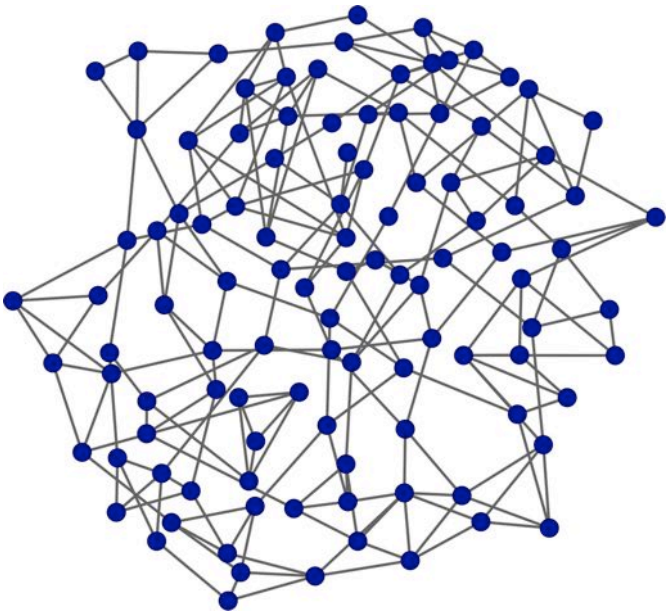


Figura 23-6: Ilustração de uma rede aleatória consistindo em 109 proteínas. A rede apresenta $P(k)$ 3,8. Observe que as conexões de cada nó são valores próximos a 4, o que está de acordo com $k \approx \langle k \rangle$.

probabilidade p, gerando uma rede de conexões aleatórias com aproximadamente pN . $(N - 1) / 2$. Dessa forma, o grau dos nós segue uma distribuição de Poisson com máxima em $\langle k \rangle$ e a maioria dos nós apresentando aproximadamente o mesmo número de conexões $k \approx \langle k \rangle$, com grau próximo ao da média da rede. Raramente surgem nós que apresentam mais ou menos conexões que $\langle k \rangle$. Adicionalmente, redes aleatórias apresentam a propriedade "mundo pequeno" e distribuição de grau exponencial, sendo estatisticamente homogêneas.

Rede de livre escala

O modelo de rede de livre escala foi introduzido por Barabási e Albert em 1999 onde se observa que redes complexas, como as redes de citações de artigos científicos, redes metabólicas, redes socais e a World Wide Web apresentam distribuição de grau que segue uma lei de potência $P(k) \sim k^{-\gamma}$, $\gamma > 1$. Essas redes são consideradas como livres de escala (Figura 24-6) pois a lei de potência não permite uma escala característica.

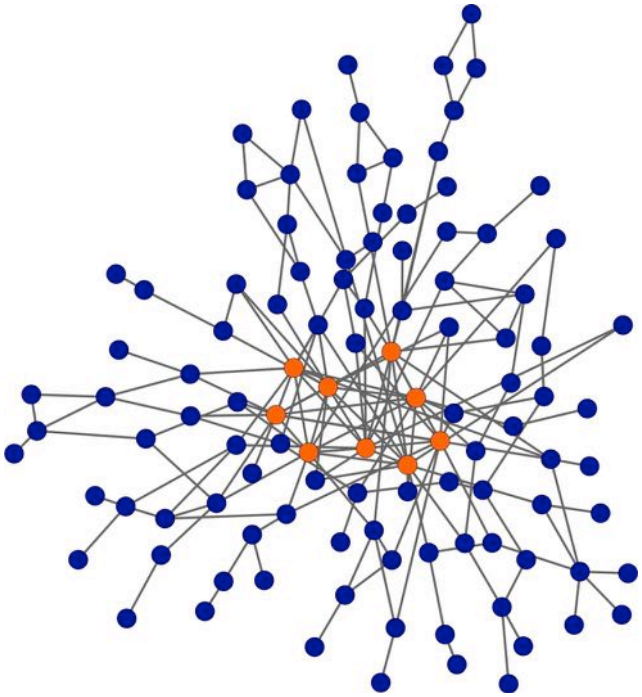


Figura 24-6: Ilustração de uma rede de livre escala consistindo de 109 proteínas, na qual o grau de distribuição segue uma lei de potência. Neste tipo de rede, as proteínas *hubs* (nós laranjas) tem papel essencial na manutenção da integridade da rede.

Diferentemente da rede aleatória que apresenta um número fixo de N nós, as redes de livre escala apresentam uma ordem dinâmica de estruturação que permite o crescimento da rede pela adição de novos nós. Assim, a rede aleatória consiste de um sistema aberto que inicia com um pequeno grupo de nós e aumenta de tamanho exponencialmente no tempo devido à inserção de novos nós. A probabilidade deste novo nó se conectar a nós com grande número de conexões é maior, sendo chamada de conexão preferencial. Por exemplo, imagine que você está buscando um artigo sobre determinado assunto na Internet. Certamente os artigos que você encontrará mais facilmente serão publicações com alto grau de conexão por serem mais conhecidos e bem citados quando comparadas a publicações pouco citadas e, consequentemente, menos conhecidas.

Estes dois mecanismos, crescimento da rede e conexão preferencial originaram o algoritmo do modelo Barabási-Albert, que estabelece que o crescimento inicia-se como uma pequena rede, sendo que a cada instante de tempo um novo nó com m conexões é adicionado, onde a probabilidade do novo nó se conectar ao nó i que está previamente presente depende de k_i (grau de i):

$$\prod(ki) = \frac{k_i}{\sum_j k_j}$$

Esse crescimento gera uma rede de livre escala com expoente de grau $\gamma = 3$. Após t instantes de tempo, temos uma rede com $N = t + m_0$ e m_t conectores.

As características da rede de livre escala a tornam uma rede que apresenta um pequeno número de nós altamente conectados (*hubs*), o que frequentemente determina suas propriedades. Como já mencionado, falhas na rede (ou remoção de nós aleatórios) apresentam poucas consequências, enquanto que o ataque aos nós altamente conectados tornará a rede fragmentada. Em sistemas biológicos, uma rede bioquímica apresenta alta resiliência contra mutações aleatórias, enquanto que os *hubs* podem ser usados como candidatos importantes para alvo de fármacos. Um exemplo disso seria a proteína EF-Tu. Esta proteína tem papel essencial durante a elongação da síntese proteica, sendo inibida pelo antibiótico quirromicina, que impede que o complexo EF-Tu-GDP seja liberado do ribossomo.

Rede Hierárquica

Como já vimos anteriormente, uma rede pode ser avaliada pelo grau de agrupamento (clusterização) de seus nós. Na maioria das redes baseadas em um sistema real (chamadas de redes reais), como por exemplo, parte de uma via metabólica, o coeficiente de clusterização é significativamente maior se comparado a redes aleatórias. Da mesma forma, ocorre a coexistência da propriedade de livre escala e clusterização nas redes reais, como redes metabólicas e de interação proteica. Contudo, grande parte dos modelos propostos para representar estas redes não consegue descrever a livre escala e a clusterização simultaneamente.

Adicionalmente, muitas redes reais

apresentam módulos, ou seja, a rede é composta de subredes funcionalmente separáveis. Esses componentes separáveis apresentam densa conectividade entre os seus próprios nós, com conectividade mais dispersa em relação a componentes de outros módulos. Isso ocorre porque cada módulo apresenta a capacidade de executar uma tarefa identificável, diferente de outro módulo. Contudo, essa "separação" de tarefas não significa que um módulo é independente de outro, mas sim que tem funções distintas.

Dessa forma, é necessário combinar a propriedade de livre escala, o alto grau de agrupamento e a modularidade de uma forma interativa, gerando a rede hierárquica. A estrutura hierárquica é convencionalmente representada por um dendrograma ou uma árvore e atua relacionando os nós mais próximos na rede, conforme Figura 25-6. Essas redes podem ser formadas basicamente pela duplicação de *clusters* e repetidas indefinidamente, integrando uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independentes do tamanho do sistema. Muitas vezes, em redes reais, a modularidade não apresenta um limite claro, sendo reconhecida principalmente por nós altamente conectados entre si e conectados a outros módulos.

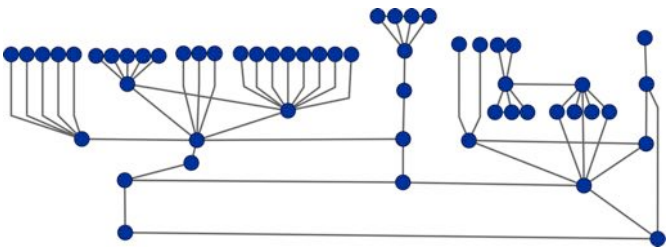


Figura 25-6: Ilustração de uma rede hierárquica consistindo de 55 proteínas em modelo de dendrograma onde é possível observar sua modularidade intrínseca.

A principal característica dessas redes que não é compartilhada por redes aleatórias ou de livre escala é a hierarquia intrínseca, sendo representada também na sua arquitetura. Essa característica hierárquica pode ser, ainda, analisada quantitativamente, como observado por Dorogovtsev e colaboradores em 2002, que construíram um gráfico de livre escala determinístico, na qual o coeficiente de clusterização de um nó que possui k conexões segue a lei de escala $C(k) \sim k^{-1}$. Portanto, o modelo de rede hierárquico integra uma topologia livre de escala com alta modularidade, resultando em um coeficiente de clusterização independente do tamanho do sistema.

6.6. Perturbação e conectores

Como visto anteriormente, um grafo consiste de um conjunto de nós e um conjunto de conectores que conectam esses nós. Portanto, os nós são as entidades de interesse e os conectores representam as relações entre as entidades.

Quando tratamos de sistemas biológicos, podemos levar em consideração diferentes entidades como, por exemplo, DNA, RNA, metabólitos, pequenas moléculas e/ou proteínas. Estes componentes biológicos não atuam isoladamente, mas sim dependem da interação com outros componentes. Para que ocorra essa interação (comunicação) é necessária a presença de conectores.

Conectores podem ser interações físicas, bioquímicas ou funcionais. Por exemplo, em redes metabólicas, conectores podem ser reações que convertem um metabólito em outro ou enzimas que catalisam essas reações; em redes de regulação gênica, conectores podem representar a ligação física de um fator de transcrição nos elementos regulatórios; em redes de doenças, conectores podem representar as mutações genéticas associadas à doença; e em redes proteicas, os conectores podem ser ligações físicas entre as proteínas.

Como apresentado anteriormente, as redes podem ser direcionadas e não direcionadas. Esse comportamento da rede depende da natureza da interação e, obviamente, da direcionalidade dos conectores (Figura 26-6). Em redes direcionadas, a interação entre dois nós tem uma direção bem definida que representa, por exemplo, a direção do fluxo do substrato ao produto em uma rede metaból-

ca. Em redes não direcionadas, a ligação não tem uma direção definida, tal como a interaçao física entre proteínas.

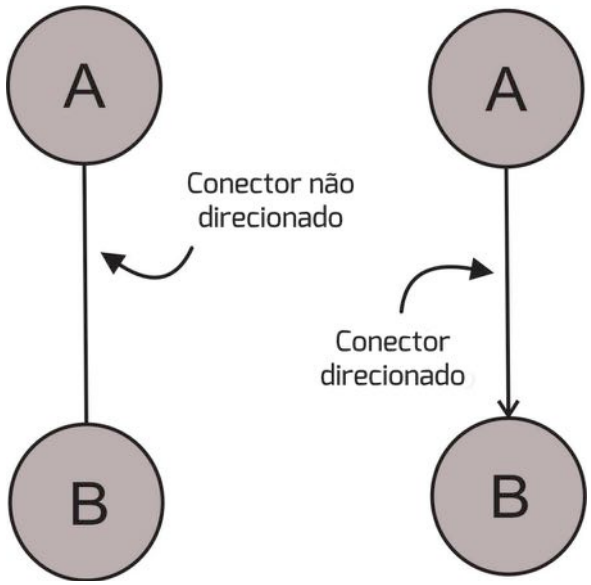


Figura 26-6: Representação de um conector não direcionado e um direcionado.

Na abordagem da biologia de sistemas tão importante quanto conhecer os nós que interagem entre si em uma rede é compreender, por exemplo, que tipo de interação pode ocorrer na rede em questão, quais conectores são mais relevantes à rede e qual o impacto da perturbação de um conector. Nesta seção iremos discutir os tipos de conectores entre diferentes componentes de uma rede envolvendo proteínas e as consequências da ruptura nestas conexões.

Interação proteína-proteína

A interação proteína-proteína é comum e crucial a vários processos celulares, tais como na ligação enzima-inibidor e na interação antígeno-anticorpo. Os diferentes tipos de complexos proteicos têm sido definidos na literatura como obrigatórios e não obrigatórios. No complexo obrigatório, as proteínas não podem funcionar separadamente, diferindo do complexo não obrigatório onde as proteínas associam-se e dissociam-se dependendo de fatores externos, podendo também exercer funções fora do complexo.

De acordo com a estabilidade e o mecanismo de formação do complexo, incluindo o tipo de conexão entre as proteínas, as interações podem ser conceitualmente separadas em dois grupos: aquelas que são permanentes e aquelas que são temporárias. E, embora não exista um limite bem definido para essa separação, tendências têm sido observadas em relação a suas propriedades biológicas (Figura 27-6).

Em relação à estrutura, por exemplo, interações temporárias são caracterizadas por interfaces proteicas pequenas, enquanto que as interfaces de proteínas interagindo permanentemente são maiores. Consequentemente, complexos proteicos com interfaces maiores tendem a apresentar um maior grau de mudança conformacional após a ligação. Além disso, componentes de complexos permanentes tendem a ser co-expressos e mais estáveis. Esta estabilidade gera uma pressão seletiva maior e em função disso, uma taxa evolutiva mais lenta.

Como será discutido adiante, interação transitória tende a ser *date*, isto é, as proteínas podem se conectar em diferentes tempos e a interação permanente tende a ser *party*, isto é, conexão proteica forte e constante.

As proteínas com conectores permanentes existem somente em sua forma complexada e são muito estáveis, enquanto aquelas com conectores transitórios possuem a capacidade de associação e dissociação *in vivo*. Dentre as proteínas com conectores transitórios, há aquelas em que a associação/dissociação é resultante de uma conexão com baixa afinidade, porém constante (interações temporárias fracas) e aquelas em que a associação/dissociação é desencadeada por um processo ativo (interações temporárias fortes) como, por exemplo, uma mudança conformacional ocorrida em consequência de um fator ligante.

A diferença entre as interações acima citadas é distinguida puramente pelas propriedades da estrutura da interface proteica, isto é, da superfície de contato das proteínas. Essas propriedades conferem afinidade e especificidade, e são determinadas principalmente por forças intermoleculares como comple-

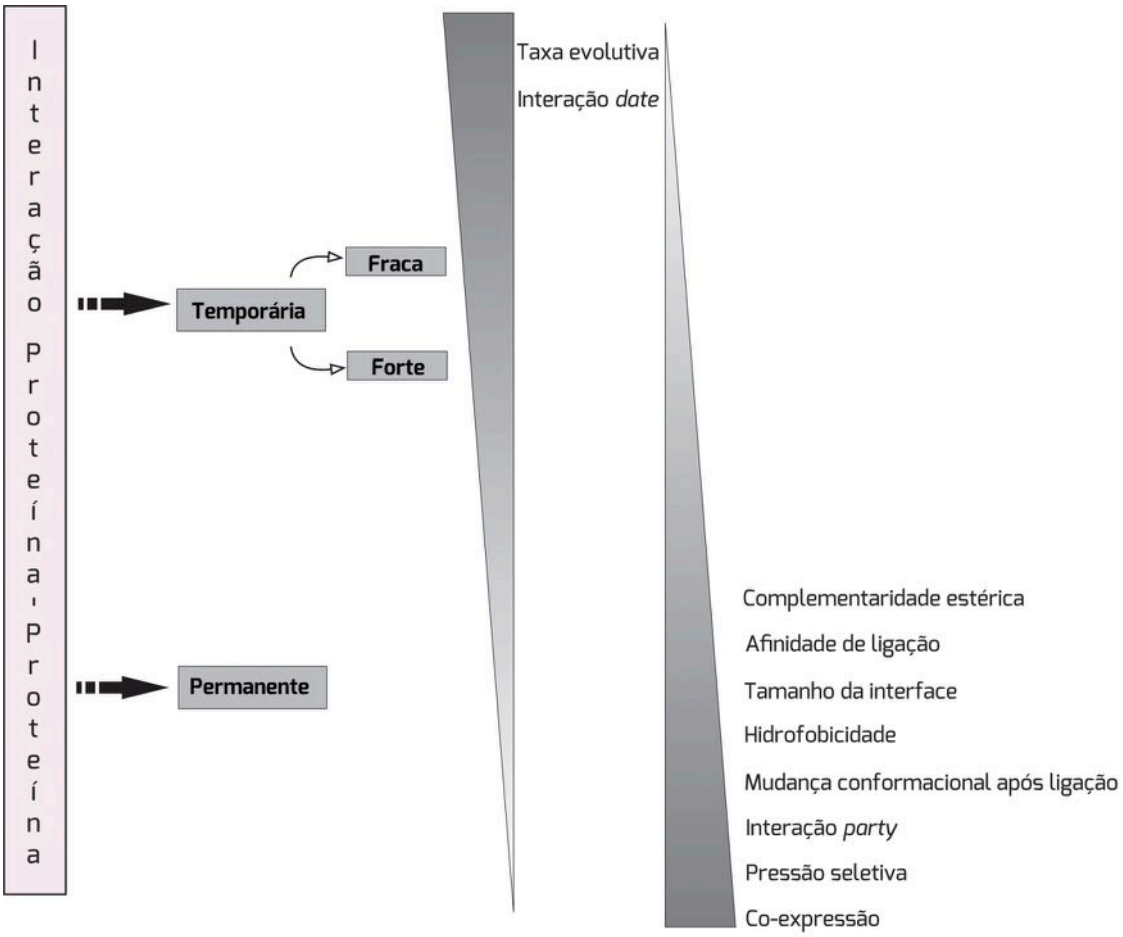


Figura 27-6: Modelo esquemático representando os diferentes tipos de interações proteína-proteína e as propriedades biológicas relacionadas. Quanto maior o tamanho da base e a intensidade da cor do triângulo, maior é a relação entre o modo de interação proteica e a propriedade biológica.

mentaridade estérica, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A complementaridade estérica otimiza as interações de van der Waals entre o complexo. Normalmente, estas interações de fraca energia ocorrem em função da polarização transiente de ligações carbono-hidrogênio ou carbono-carbono e, apesar de fracas, são extremamente importantes para o processo de reconhecimento intermolecular pois crescem em intensidade com a área de interação. Complexos com conexões permanentes exibem alta complementaridade estérica nas proteínas em contato, enquanto complexos com conexões temporárias demonstram baixa complementaridade.

Como as interações de van der Waals, as interações hidrofóbicas são pontualmente fracas e ocorrem em função da interação entre cadeias ou subunidades apolares. Os complexos com conexões permanentes normalmente persistem no estado ligado, sendo a força hidrofóbica mais significativa. Já em conectores transitórios, a alta hidrofobicidade se torna desfavorável, pois esses complexos permanecem ligados por menos tempo.

As forças de atração eletrostáticas são aquelas resultantes da interação entre dipolos e/ou íons de cargas opostas e representam força significativa na interação proteína-proteína, podendo definir o tempo de vida do complexo.

Dentre as forças intermoleculares discutidas acima, o fator dominante da interação permanente entre proteínas consiste nas interações hidrofóbicas, enquanto várias forças

participam de interações temporárias entre proteínas. Além disso, proteínas interagindo de forma temporária possuem interfaces que são menores em tamanho do que as interfaces de proteínas permanentes, os aminoácidos que compõem a interface e a proporção de resíduos hidrofóbicos não diferem drasticamente do resto da superfície proteica e as interfaces são levemente ricas em grupos polares neutros e em água.

O tipo de interação também confere graus diferentes de restrição (pressão seletiva) na evolução da proteína. Proteínas com interação permanente tendem a evoluir em uma velocidade menor comparada a proteínas que formam complexos temporários, bem como possuir pressão seletiva maior e menor plasticidade em sua sequência.

Evidências sugerem que o modelo duplicação-divergência aplica-se à evolução das redes proteicas. Uma das predições é que na duplicação das proteínas algumas ou todas as conexões podem ser herdadas da proteína ancestral. Consistente com esta hipótese, proteínas parálogas tendem a compartilhar padrões de interação em uma frequência maior do que a esperada ao acaso. No entanto, tem sido proposto que depois que a duplicação gênica ocorre, as interações entre as proteínas são rapidamente perdidas. Portanto, duplicações recentes são mais prováveis de compartilhar interações, comparadas a duplicações mais ancestrais.

Outra distinção acerca da interação proteica refere-se à interação funcional e interação física. A interação funcional pode ou não corresponder a uma interação física direta em algum processo biológico. Assim, na interação física, a proteína A conecta-se a proteína B e, na interação funcional, a proteína A atua com a proteína B. Como exemplo de interação funcional podemos imaginar dois produtos gênicos que interagem em uma mesma via em um processo biológico, mas não se conectam fisicamente.

O tipo de interação tem um papel importante na determinação do comportamento das proteínas. Como já vimos, *hubs* são proteínas envolvidas em um grande número de interações (altamente conectadas) dentro de uma rede proteica. Algumas proteínas *hub* são altamente co-expressas com outras proteínas do módulo, o que implica na existência de complexos estáveis (permanentes). Outras proteínas possuem expressão independente, sugerindo a ligação com proteínas em diferentes tempos, de modo transitório. Esses *hubs* são classificados como *party* e *date hubs*, respectivamente.

Na construção de redes proteicas, a diferenciação entre complexos permanentes e transitórios tem importantes implicações. Por exemplo, na prospecção de novos fármacos, a alteração do padrão de interação entre proteínas temporárias por modulação farmacológica ocorre mais facilmente em comparação a proteínas que formam complexos permanentes. Portanto, uma rede de interação proteica não é um processo estático, mas sim corresponde a um constante fluxo de informações. Por conseguinte, na análise de dados de interação proteína-proteína a discriminação das características da interação e/ou o uso de centralidades de conectores é fundamental para obter modelos mais realísticos.

Interação proteína-ácidos nucleicos

Proteínas que se ligam a ácidos nucleicos têm um papel central em todos os processos regulatórios que controlam o fluxo de informação genética. Por exemplo, proteínas podem inibir, ativar e coordenar a transcrição do DNA, auxiliar e manter o empacotamento e o rearranjo do DNA e o processamento do RNA, coordenar a replicação do DNA, promover a síntese de proteínas e sinalizar o reparo do DNA, entre outros.

Esses possíveis papéis fisiológicos são determinados pela afinidade e especificidade da interação DNA-proteína, que é a habilidade da proteína em distinguir seu sítio de ligação do restante do DNA. Estas propriedades dependem de interações precisas entre a sequência de aminoácidos da proteína e os nucleotídeos do sítio específico de ligação do DNA.

As proteínas que se ligam a ácidos nucleicos podem ser, de forma simplificada separadas em três grupos de acordo com a função:

i) enzimas, onde a principal função da proteína é modificar a organização do ácido nucleico, como no caso das endonucleases, glicosiltransferases, glicosilases, helicases, ligases, metiltransferases, nucleases, polimerases, recombinases, topoisomerases, translocases e transposases, entre outras;

ii) fatores de transcrição, onde a principal função da proteína é regular a transcrição e a expressão gênica como por exemplo, TFIIA, TFIIB, TFB, entre outros;

iii) proteínas estruturais que ligam-se ao DNA, que têm como principal função suportar a estrutura e a flexibilidade do DNA ou agregar outras proteínas, por exemplo, proteínas centroméricas, proteínas envolvidas no empacotamento e na manutenção/proteção do DNA, proteínas de reparo, proteína envolvidas na replicação e proteínas teloméricas, entre outras.

A interação proteína-proteína também é necessária para uma eficiente interação entre proteínas e ácidos nucleicos. A interação proteína-proteína com o DNA pode ocorrer de três modos de acordo com a direção e o eixo da dupla hélice do DNA (Figura 28-6):

i) a direção da interação entre as proteínas e o eixo da dupla hélice é perpendicular;

ii) a direção da interação da proteína é paralela ao eixo da dupla hélice;

iii) ambos os modos de interação são observados ao mesmo tempo.

Assim como na formação de complexos proteicos, discutido anteriormente, a formação de complexos DNA-proteína ou RNA-proteína também envolve forças intermoleculares, tais como van der Walls, força eletrostática, interação hidrofóbica e ligações de hidrogênio.

A região da proteína que reconhece a sequência do ácido nucleico é denominada motivo. Os motivos hélice-volta-hélice, dedo de zinco e zíper de leucina são os mais comuns encontrados nas proteínas que interagem com ácidos nucleicos.

O motivo hélice-volta-hélice é um dos elementos normalmente encontrados nos fatores de transcrição e nas enzimas de procariotos e eucariotos, sendo formado por duas hélices α conectadas por uma volta. O motivo liga-se a cavidade maior do DNA e, em muitos complexos, o contato direto é feito entre a cadeia de aminoácido e a sequência de bases do ácido nucleico.

Já o motivo dedo de zinco é encontrado principalmente em fatores de transcrição de eucariotos. Um dedo de zinco é composto por duas folhas β antiparalelas e uma hélice α, sendo o íon zinco fundamental para garantir a estabilidade deste tipo de domínio. Subunidades proteicas contêm múltiplos dedos de zin-

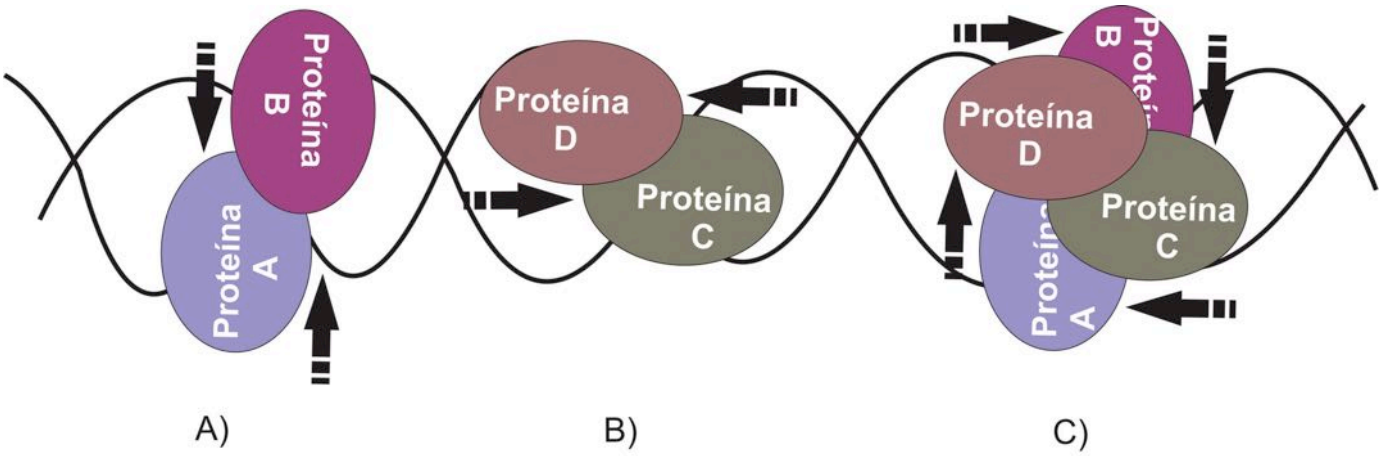


Figura 28-6: Modos de interação proteína-proteína com a dupla hélice do DNA. A) perpendicular; B) paralela e C) ambas as direções são observadas.

co que se enrolam no DNA formando uma espiral, inserindo a hélice α na cavidade maior do DNA.

Fatores de transcrição de eucariotos e procariotos também podem conter o motivo zíper de leucina, encontrado em proteínas regulatórias. Esse motivo é formado por duas hélices α paralelas, unidas por resíduos de leucina.

A estrutura do zíper de leucina pode ser dividida em duas partes: a região de dimerização e a região de ligação ao DNA. A dimerização é mediada pela formação de uma estrutura enrolada na região carboxi-terminal de cada hélice com sete resíduos de leucina. A região que se liga ao DNA, também conhecida como região básica, é encontrada na região amino-terminal da hélice que se projeta na cavidade maior do DNA. Embora motivos de diferentes famílias de DNA sejam similares estruturalmente, pouca homologia é observada fora do motivo. Há baixa identidade entre motivos de diferentes famílias de proteínas e esta variação permite, portanto, o reconhecimento de diferentes conjuntos de sequências de DNA. Além disso, a posição do domínio dentro da cavidade maior do DNA também varia, refletindo a necessidade funcional e estrutural de cada proteína.

A afinidade e a especificidade na ligação de proteínas ao DNA não podem ser endereçados somente a alguns resíduos de aminoácidos, mas o envolvimento de toda a proteína deve ser considerado. Por exemplo, a maioria das proteínas que se ligam ao DNA possuem domínios desordenados que contribuem para o reconhecimento do DNA em vários níveis.

Proteínas com domínios desordenados são proteínas que não apresentam estrutura 2^{ária} e 3^{ária} sob condições fisiológicas e na ausência de ligantes naturais. Essas proteínas possuem alta especificidade e baixa afinidade na interação, são capazes de interagir com mais de uma proteína e alvos de modificações pós-traducionais, possuindo a capacidade de manter sua função mesmo em ambientes extremos. Na interação com o DNA, o domínio desordenado da proteína não é crucial à formação do complexo, mas pode influenciar o reconhecimento da sequência do DNA, conferindo seletividade e afinidade de ligação.

Além da característica das cavidades na molécula de DNA, da presença de motivos específicos nas proteínas ou ainda da ocorrência de domínios desordenados, outros fatores podem influenciar a interação do DNA-proteína, tais como a flexibilidade e a afinidade da proteína pelo DNA e presença de água no meio.

Muitas proteínas são flexíveis ao ponto de alterar sua conformação quando se ligam ao DNA, enquanto outras são conhecidas por alterar a conformação do DNA após a ligação. A afinidade da interação entre o DNA e uma proteína tende a estar relacionada à relevância funcional da proteína. Por exemplo, a afinidade de um fator de transcrição por seu sítio de ligação é proporcional à ativação que ele exerce. Ainda, alguns contatos mediados por água foram observados entre proteínas e o DNA, participando de redes de ligações de hidrogênio que conferem estabilidade ao complexo.

Interação entre proteínas e pequenos compostos

Considerando-se que a interação proteína-proteína normalmente envolve superfícies relativamente grandes, pode-se imaginar que moléculas menores não seriam efetivas na modulação da ligação dos complexos por apresentarem áreas menores e, por conseguinte, interações menos intensas. Contudo, ao empregarmos estruturas químicas diferentes de aminoácidos, podemos não só compensar esta redução na área de contato mas produzir moléculas com afinidade maior do que os próprios ligantes fisiológicos envolvidos do processo de interesse.

Adicionalmente, estas moléculas de baixa massa molecular tendem a apresentar muitas vantagens terapêuticas em relação a proteínas, dentre as quais se destaca sua maior estabilidade metabólica e consequente maior biodisponibilidade. Podem atuar diretamente – via inibição da interface proteína-proteína – ou indiretamente – via ligação a um sítio alostérico que induz uma mudança conformacional do alvo da proteína ou da molécula associada.

A busca de novos fármacos deve levar em conta o tipo de complexo proteico alvo. A formação de complexos permanentes pode ser considerada uma continuação do enovelamento da proteína, sendo o dobramento fi-

nal das subunidades parte deste processo. Assim, esse tipo de complexo é menos propenso à modulação farmacológica, sendo mais interessante explorar o processo de dobramento em si como alvo de pequenos compostos. Já as interfaces das proteínas de complexos temporários são alvos efetivos ao planejamento de novos moduladores terapêuticos.

Para que pequenas moléculas modulem a interação proteica, estratégias têm sido estabelecidas e dois principais mecanismos do controle regulatório têm sido utilizados: a inibição e a estabilização (Figura 29-6). Das estratégias mais exploradas, destaca-se a inibição da interação proteína-proteína.

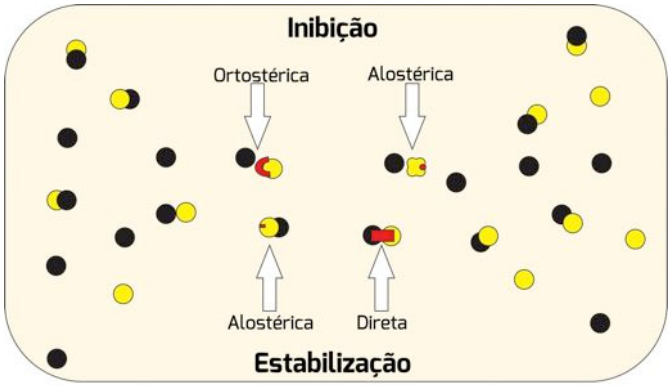


Figura 29-6: Dois principais mecanismos de modulação da interação proteína-proteína utilizando pequenos compostos. Diferentes proteínas são apresentadas em preto e amarelo. Pequenos compostos são apresentados em vermelho.

O modo de ação da maioria dos inibidores de interação proteica é baseado na ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos *hot spots* críticos da interface e competindo com a proteína original. Esse tipo de inibição é conhecido como ortostérica. Na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante (Figura 29-6).

Pequenas moléculas estabilizadoras da interação proteína-proteína também demonstram dois modos gerais de ação. Primeiro, um estabilizador pode ligar-se a uma única proteína, na qual aumenta a afinidade de ligação mútua das proteínas do complexo de um modo alostérico. Segundo, a molécula estabilizadora liga-se à superfície do complexo proteico, fazendo contato com ambas as proteínas ligantes e aumentando a afinidade de ligação mútua entre elas. Assim, a inibição estabilizadora pode ser denominada alostérica (ligada a uma proteína) ou direta (ligada ao menos a duas proteínas).

A ativação por pequenos compostos é, normalmente, um processo mais intrincado pois, além da ligação, é necessário o correto desencadeamento da cascata de ativação. Compostos que induzem a interação proteica são chamados de dimerizadores. Inúmeras vias de sinalização celular iniciam a partir da dimerização proteína-proteína. A principal ideia do uso de dimerizadores é a indução de interação entre duas proteínas por pequenas moléculas que levam à ativação da via de sinalização celular. Na literatura científica foi observado que dimerizadores podem induzir proliferação celular, transcrição e apoptose.

Perturbação dos conectores

Perturbações podem ocorrer em todos os sistemas, e em sistemas biológicos não é diferente. Nos interatomas, essas perturbações podem variar desde a remoção de um ou mais nós até a remoção de conectores. Desta forma, as consequências na estrutura e na função do sistema irão diferir drasticamente dependendo do tipo de perturbação ao qual a rede foi exposta. Como exemplo, podemos imaginar uma rede de proteínas que confere um fenótipo específico (Figura 30-6).

A remoção do nó não somente incapacita a função deste, mas também a de outros nós, causando a ruptura nas vias de todos os nós vizinhos. Uma perturbação no conector, que remove uma ou poucas interações mas deixa o restante da rede intacta e funcionando, pode ter efeitos mais sutis no sistema, não necessariamente alterando o fenótipo. Contudo, a consequência do desarranjo da rede após a remoção de nós ou de conectores depende da importância do nó e do conector à rede. Essas informações de conectores e nós

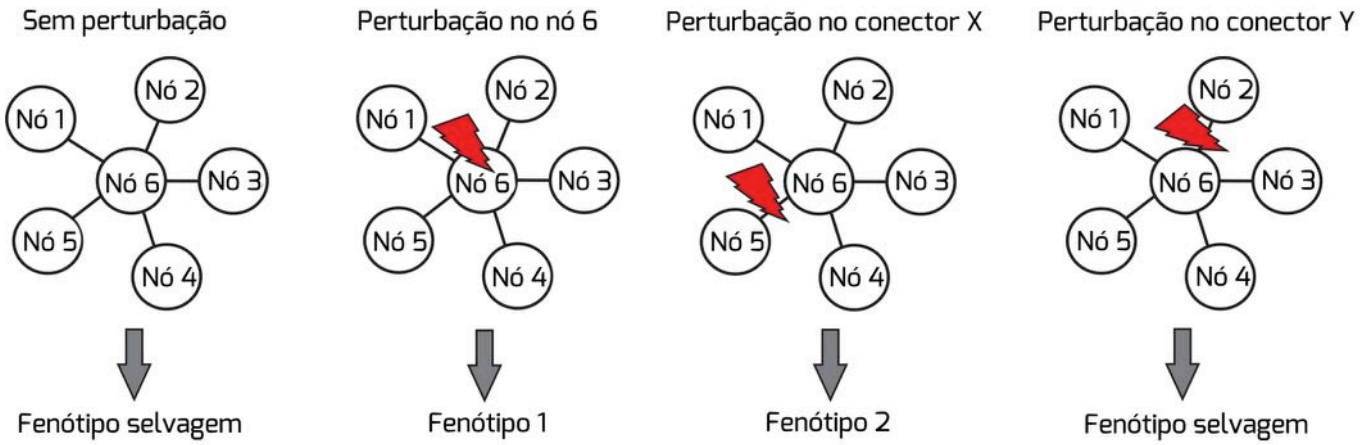


Figura 30-6: Rede hipotética de proteínas relacionada a um fenótipo específico representando diferentes tipos de perturbação e suas consequências. Neste exemplo o nó 5 e o conector entre os nós 5 e 1 são essenciais à manutenção do fenótipo selvagem.

mais informativos de uma rede podem ser obtidas, por exemplo, pela análise da resiliência e percolação da rede, vista anteriormente.

A distinção entre modelos de remoção de nó e perturbação de conectores - alteração interação-específica e conector-específica (*edge-specific* ou "*edgetic*"), respectivamente - pode providenciar novas pistas nos mecanismos básicos de doenças humanas, tais como diferentes classes de mutações que levariam a modos dominantes ou recessivos de herança genética.

Em uma rede proteica, a remoção de um nó pode representar a remoção de uma proteína, causado por uma mutação crítica no gene que desestabiliza a estrutura da proteína. Já a remoção de um conector pode representar uma mudança específica em distintas interações bioquímicas e biofísicas, preservando certos domínios da proteína.

Em relação a genes envolvidos em múltiplas doenças, foi demonstrado que alelos *edgetic* responsáveis por diferentes doenças consistem em distintas perturbações *edgetic* que, por sua vez, tendem a estar localizados em diferentes domínios de interação proteica, conferindo fenótipos diferenciados.

Pesquisadores analisaram cerca de 50.000 alelos mendelianos associados a doenças genéticas hereditárias e observaram que aproximadamente a metade foi potencialmente *edgetic*. Nesta análise foram consideradas deleções e mutações truncadas dentro dos domínios da proteína que grosseiramente desestabilizaram a estrutura da proteína, como remoção de nó, mutações com alteração em quadro de leitura que afetaram sítios de ligação específicos e mutações truncadas que preservaram certos domínios da proteína como perturbação *edgetic*. Alelos truncados foram menos propensos a expressar proteínas estáveis em comparação a alelos que alteraram o quadro de leitura, podendo diferir doenças hereditárias mendelianas envolvendo remoção de nó *versus* perturbação *edgetic*.

Um alelo *edgetic* pode ser identificado pela falta de um subconjunto de interações, quando possuem defeitos nas interações provavelmente devido a mudanças específicas dentro ou próximo a sítios de ligação da proteína ou quando fenótipos *in vivo* diferem daqueles causados por perturbações nulas (genótipos nulos).

Dependendo da rede, o fenômeno de perturbação de um único conector pode ser mais provável do que da remoção de um nó. Dependendo do conector rompido, o impacto à rede pode ser maior, pois diferentes conectores (interações) têm diferentes níveis de importância (vulnerabilidade). Conectores com alto valor de *edgebetweenness* podem causar fragmentação da rede em componentes desconectados, caso sejam rompidos, como por exemplo no caso de conectores entre *clusters*. Esse tipo de conector é assim chamado de *cut-edge*. Já conectores com baixo valor de *edgebetweenness*, quando eliminados da rede, podem ser substituídos por vias alternativas, como por exemplo no caso de

conectores dentro de *clusters*. Assim, conectores *interclusters* tendem a ser mais vulneráveis quando comparados aos conectores *intraclusters* em uma determinada rede.

6.7. Conceitos-chave

Assortatividade: tendência de nós interagirem com nós similares a eles mesmos.

Betweenness: parâmetro que estima a relação entre dois nós, ou seja, leva em consideração a quantidade de caminhos mais curtos que passam entre eles.

Biologia de sistemas: área da bioinformática que estuda sistemas moleculares complexos e como as moléculas interagem entre si.

Caminho: sequência consecutiva de nós em um grafo sem repetições, estando cada nó adjacente interligado por um conector.

Caminho geodésico: definido pela via mais curta dentro de uma rede entre dois nós quaisquer.

Circuito: sequência de nós sem repetição com um conector entre cada par de nós adjacentes na sequência, onde o nó inicial coincide com o nó final.

Clique: é definido como um grafo com alta conectividade entre seus elementos integrantes. Sendo assim, clique também é considerado um sinônimo de *cluster*.

Closeness: valor que indica os caminhos mais curtos entre um nó n e todos os outros nós da rede, uma tendência de aproximação ou isolamento de um nó.

Complexo proteico: grupo de proteínas formado pela associação de duas ou mais cadeias polipeptídicas.

Comprimento do caminho: definido pelo número de conectores que definem o caminho, ou então, pelo número de nós da sequência menos um.

Conector *Cut-edge*: conector que quando rompido causa fragmentação da rede.

Date hubs: são hubs que se ligam a diferentes proteínas em diferentes módulos (intermódulo), ou seja, diferente tempo e/ou espaço, consequentemente, apresentado um papel global na rede.

Desassortatividade: tendência de nós interagirem com nós diferentes deles mesmos.

Diâmetro: indica a distância entre os dois nós mais afastados entre si de uma rede. Sendo assim, definimos que uma rede possui um alto diâmetro quando a distância geral entre os nós é muito ampla. Quando a distância entre os nós é pequena, então o diâmetro é baixo.

Dimerização: corresponde à união de dois monômeros, formando um dímero. Ou seja, é a formação de uma molécula a partir de duas moléculas menores.

Dimerizadores: compostos que induzem a dimerização, neste caso a interação proteica.

Distribuição de Poisson: distribuição aplicada a probabilidade de ocorrência de um evento em determinado intervalo de tempo.

Edgebetweenness: parâmetro que indica o número de caminhos mais curtos entre pares de nós que percorrem um determinado conector.

Edgetic: perturbação causada em um conector específico, portanto em uma interação específica na rede.

Forças intermoleculares: forças que mantêm as moléculas unidas durante a interação.

Gargalo (*bottleneck*): proteína que apresenta alto grau de *betweenness*.

Grau de nó (*node degree*): parâmetro referente à quantidade de nós adjacentes (diretamente conectados) a outro determinado nó.

Hipergrafo: rede caracterizada pela presença de hipervértices.

Hipervértices: Conectores que interligam nós que apresentam propriedades distintas nos hipergrafos.

Hot spot proteico: locais essenciais da interface com alta afinidade de ligação.

Inibição alostérica de uma proteína: na inibição alostérica, pequenos compostos ligam-se a sítios diferentes, causando mudança conformacional suficiente para interferir na ligação da proteína ligante.

Inibição ortostérica de uma proteína: inibição causada pela ligação direta de uma pequena molécula à superfície de interação da proteína ligante, interferindo diretamente nos *hot spots* críticos da interface e competindo com a proteína original.

Interface proteica: área através da qual as macromoléculas se comunicam e exercem sua funcionalidade.

Modularidade (clusterização): padrões de conectividade, onde seus elementos constituintes estão agrupados em subconjuntos altamente conectados.

Multiconector, interações: quando há dois ou mais conectores ligando os mesmos nós na rede em redes direcionadas.

Multidígrafo: rede direcionada com a presença de multiconectores.

"Mundo pequeno", efeito: define que existe um caminho mínimo entre um nó de origem e um nó de destino.

Ontologia gênica: tipo de análise que tem como função, em uma rede de interação proteína-proteína, agrupar proteínas que façam parte de um mesmo processo biológico.

Party hubs: proteínas altamente ligadas dentro do seu próprio módulo (intra-módulo), ou seja, ligação no mesmo tempo e/ou espaço.

Pleiotrópico, efeito: proteínas pleiotrópicas são aquelas que apresentam múltiplos efeitos em um sistemas biológico.

Rede: representação gráfica da interação entre nós por meio de vértices.

Rede bipartida: existe uma partição da rede, por exemplo, partição A e partição B, sendo os nós presentes na partição A adjacentes apenas a nós da partição B, e vice-versa.

Rede direcionada: apresentam conectores que orientam o fluxo da informação em uma direção.

Rede não direcionada: os conectores desta rede não apresentam uma direção orientada.

Rede ponderada: são redes que se caracterizam pela presença de atributos associados a conectores e nós.

Resiliência: capacidade de uma rede a tolerar a deleção de seus nós por falha ou ataque.

Taxa evolutiva: medida das mudanças ocorridas numa entidade (gene, proteína, organismo, população) evolutiva ao longo do tempo.

Teoria da Percolação: tem por objetivo investigar o comportamento das propriedades de conectividade de uma rede.

Topologia de redes: estrutura e disposição de conexões entre os nós.

Vulnerabilidade do conector: grau de importância do conector.

6.8. Leitura recomendada

BARABÁSI, Albert-László; OLTVAI, Zoltán N. Network biology: understanding the cell's functional organization. ***Nat. Rev. Genetics***. 5, 101-113, 2004.

GURSOY, Attila; KESKIN, Ozlem; NUSSINOV, Ruth. Topological Properties of Protein Interaction Networks from a Structural Perspective. ***Biochem. Soc. Trans***. 36, 1398-1403, 2008.

LEVY, Emmanuel D.; PEREIRA-LEAL, Jose B. Evolution and Dynamics of Protein Interactions and Networks. ***Cur. Op. Struct. Biol.*** 18, 1–9, 2008.

MASON, Oliver; VERWOERD, Mark. Graph theory and networks in Biology. ***IET Systems Biol***. 1, 89-119, 2007.

NEWMAN, Mark E. J. The structure and function of complex networks. ***SIAM Rev***. 45, 167-256, 2003.

YU, Haiyuan; et al. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. ***PLoS Comp. Biol***. 3, e59, 2007.

WAGNER, Günter P.; PAVLICEV, Mihaela; CHEVERUD, James M. The road to modularity. ***Nat. Rev. Genetics***. 12, 921-931, 2007.