REPRESENTAÇÃO DE CÉLULAS COMPLETAS UTILIZANDO REDES COMPLEXAS

PAULO EDUARDO PINTO BURKE

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Paulo Eduardo Pinto Burke: *Representação de Células Completas Utilizando Redes Complexas* , Dissertação de Mestrado apresentada ao Instituto de Ciência e Tecnologia - UNIFESP, como parte das atividades para obtenção do título de Mestre em Ciência da Computação, Orientador: Prof. Dr. Marcos Gonçalves Quiles, Fevereiro 2016

RESUMO

A modelagem de sistemas biológicos moleculares em formato de redes vem crescendo ao decorrer dos anos possibilitando visões cada vez mais amplas de sistemas celulares. Esta disponibilidade de dados deu início à migração de estudos reducionistas para estudos com um ponto de vista mais sistêmico com abordagens *top-down*. O evidente crescimento em tamanho das redes modeladas nesse campo de pesquisa abriu portas para ferramentas matemáticas no auxílio de seu entendimento. Diversas ferramentas da teoria de redes complexas são hoje aplicadas na caracterização e obtenção de informações que se referem ao comportamento do sistema como um todo. Mesmo com estes avanços no sentido de uma filosofia de estudo mais sistêmica, redes biológicas modelam sistemas celulares sob diferentes perspectivas não havendo atualmente um sistema de modelagem que permita a integração do objeto de estudo de todos estes modelos. Este trabalho tem como objetivo entender os atuais modelos de redes biológicas, estudar as ferramentas de redes complexas aplicadas ao entendimento de sistemas de grande porte e propor um sistema de modelagem que possibilite a integração de dados biológicos em um único modelo de rede. Como estudo de caso, será modelada e analisada uma rede no formato proposto nesse trabalho sobre o organismo *Mycoplasma genitalium* o qual é uma bactéria muito estudada atualmente por sua simplicidade e também por haver interesse médico uma vez que é um patógeno humano.

ABSTRACT

The modelling of biological systems in a network format is growing in use in the past years conducting to more extensive insights to cellular systems. The big amount of data availableness started a change of point of view from a reductionist to a systemic approach. The size growing of this kind of networks opens doors to mathematical tool in order to collaborates with its understanding. A wide range of tools provided by complex networks theory are applied to help the understanding and characterization of systems as a whole. Regardless the recent migration to a systemic approach, biological networks are still modelling different aspects of living systems as we miss an integrative modelling of the object of study of those biological network models. The goal of this work is to study the current network models as well the complex network tools applied to enhance the understanding of these large models and propose a new method to integrate this kind of data in a single comprehensive network model. As a

study case, a network such as the proposed in this work will be modelled and analysed regarding the *Mycoplasma genitalium* organism, which is a well studied bacterium for its simplicity and have a significant medical relevance once it is a human pathogen.

"Do or do not. There is no try."[1]

— Yoda

AGRADECIMENTOS

Agradeço primeiramente e imensamente a minha falecida Tia Neide, sem a qual eu nunca teria chegado onde estou nem seria a pessoa que sou. Mãe, não fique com ciúme, agradeço a você por todo o suporte e incentivo (mesmo dizendo para eu arranjar um emprego de verdade). Agradeço a meu Pai que nunca me deixou desistir de meus sonhos. Aos meus orientadores Quiles e Cláudia que acreditaram em mim e me deram todo o suporte de que necessitei. Aos meus companheiros de laboratório Jeferson e Aruã pelas críticas e boas conversas. Aos professores André Zelanis e Elisa Esposito que me mostraram como é ser grande e humilde ao mesmo tempo. E a todos que cruzaram meu caminho até hoje, de certa forma contribuíram para esse trabalho ser realizado.

Só não esquecendo de agradecer a todos que produziram café durante esse tempo.

1 Tradução do Autor: "Faça ou não faça. Não existe o tentar."

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

ACRÔNIMOS

RNA Ácido Ribonucléico

DNA Ácido Desoxiribunicléico

ATP Adenosina Trifosfato

Parte I

APRESENTAÇÃO E FUNDAMENTAÇÃO DO PROBLEMA

1 INTRODUÇÃO

O processo de estudo de um dado sistema, por muitos séculos, baseou-se em: dividí-lo em suas menores partes; compreender as partes; e enfim, compreender o todo[1]. Esta metodologia reducionista, postulada e difundida por René Descartes no século XVII permeou a abordagem científica em diversos campos de estudo. O reducionismo serviu de base para o desenvolvimento de diversos campos da ciência começando pela física, influenciando o desenvolvimento da mecânica clássica de Newton, até a concepção atomística da sociedade de Locke. No que se diz respeito às ciências biológicas, o método cartesiano desempenhou um importante papel na busca pela menor parte que compunha um ser vivo[2]. Robert Hooke no século XVII observou que uma fatia de cortiça era composta por diversas estruturas poliédricas similares nomeando-as células, do latim *cella* (pequena cavidade). Um século mais tarde (1838), Mathias Schleiden e Theodor Schwann desenvolveram a teoria celular onde assumem que todo ser vivo é composto por uma ou mais partes microscópicas chamadas células. As células foram por muito tempo o objeto mínimo de estudo da biologia, contudo, mais tarde observou-se que elas eram constituídas de partes ainda menores. A metodologia reducionista se aplicou novamente abrindo espaço para a biologia molecular.

Este novo mundo intracelular composto de proteínas, RNAs[1], DNAs[2] e outras biomoléculas gerou avanços significativos no entendimento da vida, da evolução Darwiniana e das práticas medicinais. O século XX foi tomado por experimentalistas caracterizando físico-quimicamente estas biomoléculas e desvendando suas funções. Uma extensa gama de moléculas já foi (e ainda é) catalogada e hoje se encontram disponíveis em diversos bancos de dados com fácil acesso para a comunidade científica. Tecnologias recentes chamadas *high-throughput* geram quantidades imensas de dados sobre biomoléculas como composição de proteínas, sequenciamento de DNA e RNA, estrutura atômica molecular e avançam no sentido de gerar uma completude de conhecimento no que se diz às partes que compõem uma célula.

Contudo, a metodologia cartesiana encontrou suas limitações nas ultimas décadas. No campo da física viu-se que a mecânica newtoniana já não se aplicava a sistemas em escalas extremas como subatômicas ou cósmicas sendo necessárias outras abordagens como a mecânica quântica e a teoria da relatividade. Na biologia, a limitação do reducionismo foi encontrada quando as partes, ou as biomoléculas,

1 Ácidos Ribonucléicos
2 Ácidos Desoxiribonucléicos

por si só não conseguiam explicar o comportamento complexo do todo, ou comportamento de células e organismos como observado pelo biólogo Paul Weiss:

> "Podemos afirmar definitivamente... com base em investigações estritamente empíricas, que a pura e simples inversão de nossa anterior dissecação analítica do universo, procedendo-se à reunião de todas as suas peças, seja na realidade ou apenas em nossa mente, não pode levar a uma explicação completa do comportamento nem sequer do mais elementar sistema vivo"[3, p. 267].

Mais do que entender o funcionamento de cada biomolécula, viu-se importante conhecer como e com quem elas se relacionam mapeando suas interrelações em busca de uma visão mais ampla do sistema [4]. Abordagens sistêmicas dentro da biologia molecular começaram a surgir em meados da década de 90 e anos 2000, como estudos alavancados por Hiraoki Kitano, nomeando de Biologia de Sistemas este novo campo de estudo[5]. Recorrer a abordagens mais holísticas significa também lidar com um número maior de dados. Para tal tarefa, biólogos recorreram à ferramentas matemáticas e computacionais para resolver problemas de maior escala na modelagem de sistemas vivos. Hoje em dia, pode-se considerar virtualmente impossível obter grandes avanços dentro da biologia sem um forte braço computacional.

O mapeamento de interações moleculares deu início a uma gama de redes biológicas as quais visam representar diferentes aspectos de subsistemas celulares. O metabolismo de organismos, por exemplo, são comumente representados por redes metabólicas onde enzimas e substratos são conectados por reações bioquímicas e a análise dessas redes levam a um melhor entendimento de como as células processam seus nutrientes. Para acompanhar o crescimento em tamanho e quantidade dessas redes, métodos mais abrangentes se mostraram necessários para uma melhor compreensão desses dados.

A teoria de redes complexas visa estudar características que emergem de sistemas as quais não podem ser observadas em suas partes separadamente. Redes complexas são representadas por grafos, vértices conectados por arestas, podendo representar qualquer sistema discreto. Muitas informações podem ser extraídas de sua estrutura como topologia, robustez do sistema, conectividade, comunidades, etc. Entre as diversas topologias estudadas, como os *random graphs* teorizados e demonstrados inicialmente por Erdos e Rényi [6], em muitos sistemas reais tem-se observado a presença de uma topologia chamada livre-de-escala (*scale-free*) [7] e suas propriedades promovem avanços em diversas áreas como controle de doenças e estratégias de marketing. Este campo de estudos, as redes complexas, muito contribui para o entendimento de sistemas biológicos, uma vez que podem ser considerados intrinsecamente complexos.

Ainda que se tenha dado um importante passo em direção a uma abordagem mais holística de sistemas biológicos, podemos ainda sentir evidências de um arcabouço reducionista quando subsistemas celulares são mapeados separadamente em modelos de redes distintos. Algum progresso já se iniciou em relação à uma integração desses subsistemas em trabalhos envolvendo mais de um subsistema celular como observado em [8], integrando redes metabólicas, redes de regulação transcricional e redes de sinalização, e em [9], integrando redes metabólicas e de sinalização. Contudo, ainda há muito a ser feito no sentido de se abordar sistemas biológicos de forma mais sistêmica e integrativa, evidenciando assim a necessidade do desenvolvimento de novos modelos e técnicas computacionais para lidar com estas novas abordagens e conceber um melhor entendimento sobre organismos vivos.

MOTIVAÇÃO

A modelagem de processos bioquímicos ou biológicos em formato de redes vem obtendo sucesso em diversos campos da biologia molecular possibilitando a integração de dados bioquímicos e obtendo visões mais amplas dos sistemas estudados. Entre os modelos de redes mais conhecidos, podemos destacar as redes metabólicas, redes de interação proteína-proteína, redes gênicas e redes de sinalização molecular[10]. Ao passo que mais dados são gerados para alimentar essas redes, elas caminham na direção de poder representar integralmente sistemas biomoleculares. Redes próximas a tal porte são chamadas de interactomas podendo representar modelos de quaisquer redes citadas anteriormente, porém, mais comumente se referem a redes de interação proteína-proteína.

Ferramentas de redes complexas vem sendo amplamente aplicadas na análise de interactomas para compreender suas estruturas e observar características emergentes do sistema[11]. Observa-se que diversos interactomas compartilham a mesma topologia, em outras palavras, possuem um padrão de conexões semelhante entre os nós da rede chamado de livre-de-escala[12, 13, 14], caracterizado por ter muitos nós com baixo grau, ou baixo número de conexões, e poucos nós com alto grau, ou alto número de conexões[7]. Esta topologia confere ao sistema características como robustez e resiliência (ver 2.1.2) onde essas, por sua vez, são características inerentes à vida[15]. Outras ferramentas de redes complexas aplicadas a redes biológicas, como medidas de centralidade, tem também obtido sucesso na elucidação de doenças[16] e na busca por novos alvos para antibióticos[17].

Outro tópico importante são os recentes avanços em simulação de células completas. Com a grande quantidade de informações acerca de organismos específicos, métodos matemáticos e computacionais, antes aplicados a simulações numéricas de subsistemas celulares, po-

dem agora ser integrados para compor sistemas maiores a um nível celular completo. Um recente modelo híbrido foi capaz de prever o fenótipo de células da bactéria *Mycoplasma genitalium* baseado em dados heterogêneos compilados manualmente por uma revisão de mais de 900 artigos científicos[18]. Entretanto, esses modelos ainda se baseiam na abordagem reducionista sendo que cada subsistema é modelado e simulado separadamente.

Os atuais métodos para simulação de sistemas biológicos talvez abordem os problemas de maneira reducionista por falta de modelos mais integrativos. Ainda não há um modelo único de rede que possibilite a modelagem de dados de diferentes subsistemas integrando as redes que os compõem mesmo que elas sejam de fato todas interligadas. Modelos de redes biomoleculares mais holísticos e integrativos podem alavancar o desenvolvimento de novos métodos de simulação mais abrangentes e obter um melhor entendimento de processos celulares como um todo.

OBJETIVOS

Os objetivos desse trabalho são, baseando-se nas necessidades da modelagem de uma célula completa, estudar e propor um método para construir uma rede que possa integrar diferentes subprocessos celulares de forma homogênea. Esse método será aplicado a dados sobre a bactéria *Mycoplasma genitalium* visando gerar uma rede que modele todos seus processos celulares conhecidos. Esta rede então será submetida a análises fazendo uso de ferramentas de redes complexas caracterizando-a e a validando com base em dados da literatura e experimentais. Também será proposto um método utilizando a rede a ser gerada neste trabalho para predição de genes essenciais do organismo estudado.

2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os temas que serviram como fundamentação para o desenvolvimento do projeto proposto. Visando o objetivo de criar uma metodologia consistente para gerar redes biomoleculares que tenham a capacidade de modelar de forma homogênea todos os processos de uma célula, serão discutidos tópicos em: redes complexas, abordando modelagem e análise de redes; fundamentos da biologia molecular com o objetivo de embasar o objeto de modelagem; e por fim, estudar as redes biomoleculares mais difundidas no meio 1 com o objetivo de entender suas capacidades e deficiências.

REDES COMPLEXAS

Modelagem de Sistemas Utilizando Redes

Todo sistema composto por um número discreto de entidades onde essas entidades se relacionam pode ser modelado em forma de rede. Em alguns sistemas, a relação entre suas entidades é muito evidente, como por exemplo uma rede de computadores onde eles são ligados fisicamente por cabos; os computadores compõem os nós[1] de uma rede e os cabos indicam os links[2] entre esses nós. Em outros sistemas, a modelagem em formato de redes pode não ser tão trivial, como por exemplo uma rede social onde pessoas são os nós e suas diferentes relações formam os links entre elas.

Matematicamente, uma rede é denotada por um grafo $G = (V, E)$ onde $V = \{v | v \in \mathbb{N}\}$ é um conjunto de nós e $E = \{e | e \to P(V)\}$ um conjunto de links os quais conectam pares de nós. O número de nós de uma rede é comumente denotado por n e o número de links por m. Uma das formas mais comuns de se representar redes é utilizando uma matriz de adjacência. Esta matriz é sempre uma matriz quadrada de ordem n contendo uma linha e uma coluna correspondente a cada vértice [19, p. 109,110]. Um valor diferente de zero (normalmente 1) em uma posição da matriz indica uma conexão entre os nós correspondentes à linha e coluna. Na Fig. 1 temos uma rede genérica não-direcionada com $n = 4$ e $m = 4$. A matriz de adjacência que representa essa rede (Eq. 1) é uma matriz simétrica, uma vez que dois

1 Também chamados de "vértices"
2 Também chamados de "arestas"

nós quando ligados entre si, como os nós 1 e 2, a conexão existe tanto de 1 para 2 quanto de 2 para 1:

$$G = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{1}$$

Na ausência de conexões de um nó com ele mesmo, chamados *loops*, a diagonal principal é sempre nula.

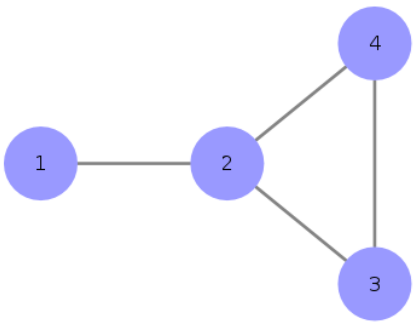


Figura 1: Rede não-direcionada com $n = 4$ e $m = 4$.

REDES DIRECIONADAS Podemos em uma rede também representar fluxo de informações entre os nós. Os links podem ser direcionados, indicando uma conexão em somente um sentido, como do nó 2 para o nó 1 na Fig. 2, não havendo conexão do nó 1 para o nó dois. A direcionalidade de uma rede implica em uma matriz de adjacência não necessariamente simétrica como na eq. 2:

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{2}$$

REDES PONDERADAS As conexões entre nós podem ter diferentes pesos, podendo representar a força ou intensidade da conexão. Na representação de uma rede na forma de matriz de adjacência, podemos substituir os valores 1 da conexão por um número qualquer desejado que indique a força dessa conexão. Na Fig. 3 a espessura das arestas

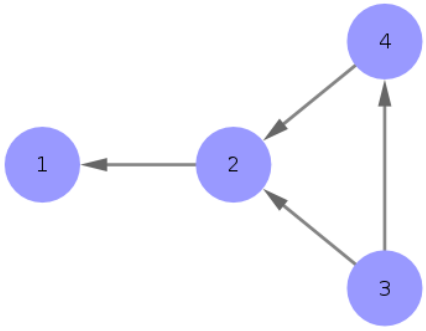


Figura 2: Rede direcionada com $n = 4$ e $m = 4$.

indicam o peso do link como indicado na sua matriz de adjacência na eq. 3.

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1.0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0.5 \\ 0 & 0.3 & 0 & 0 \end{bmatrix} \tag{3}$$

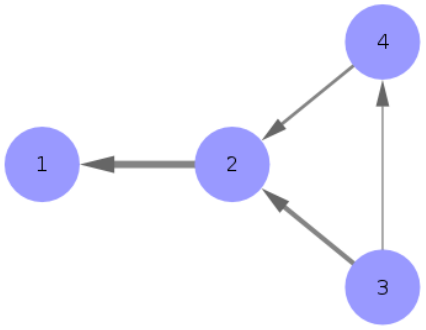


Figura 3: Rede direcionada e ponderada com $n = 4$ e $m = 4$. A espessura do link representa seu peso, quanto mais espesso, maior o peso.

REDES BIPARTIDAS Uma rede pode ser considerada bipartida quando apresentam dois conjuntos de nós disjuntos, U e V com $G = (U, V, E)$, onde não há conexões entre nós do mesmo grupo. Em um problema de coloração de grafos, ele só pode ser considerado bipartido se for bi-colorível, ou seja, se pintarmos cada grupo de vértices de uma cor, todo link apresenta cores distintas em suas terminações. Na Fig. 4 pode-se observar um grafo colorido onde o conjunto $U = \{1, 3, 4\}$ está colorido de azul e o conjunto $V = \{2\}$ está colorido de vermelho.

Uma rede bipartida pode ser transformada em uma rede comum gerando sua projeção de um modo. Para isso, é selecionado um conjunto de nós e são conectados entre si os nós que compartilham ligações em um mesmo nó do outro conjunto.

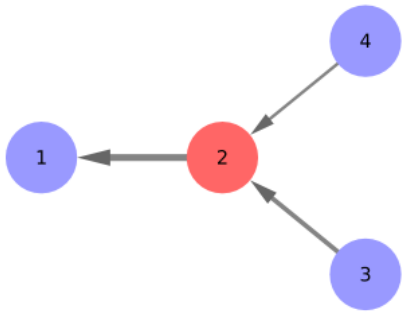


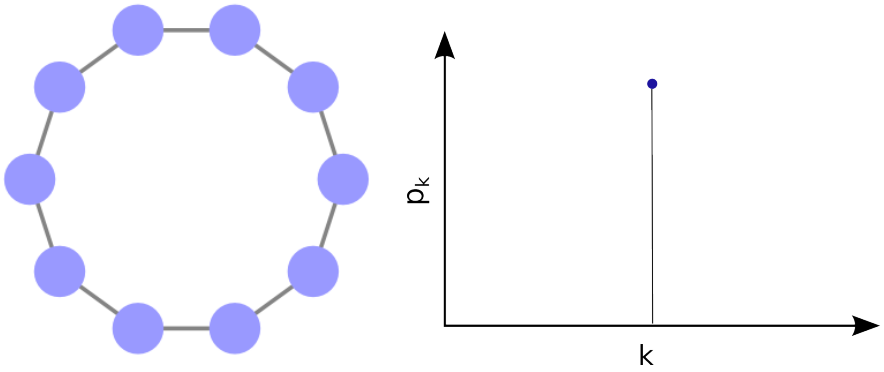
Figura 4: Rede ponderada, direcionada e bipartida com conjunto U colorido em azul e conjunto V colorido em vermelho.

Topologia

A topologia de uma rede diz respeito a como os nós de uma rede estão conectados entre si, se são ligados de forma aleatória ou seguem algum padrão, por exemplo. A principal forma de se estudar a topologia de uma rede é por meio de uma análise da distribuição de grau de seus vértices, onde o grau de um vértice quantifica o número de conexões que este vértice faz. Podemos chamar de p_k a fração de nós de uma rede com grau k, um gráfico da distribuição de grau de uma rede pode ser construído apresentando p_k em função de k. Diferentes redes podem apresentar diferentes comportamentos dessa curva, a qual, por si só, pode demonstrar propriedades de uma rede. A seguir, três topologias de rede serão apresentadas e algumas de suas propriedades discutidas.

REDES REGULARES Redes regulares, também chamadas redes lattice, são caracterizadas por ter todos os nós com um grau k constante ou com variância muito pequena. Alguns sistemas naturais e artificiais apresentam essa topologia como por exemplo a disposição das conexões entre os átomos de carbono em um diamante, com um grau $k = 4$, ou uma malha de trânsito com esquinas sendo ligadas por ruas. Na Fig. 5 pode-se observar uma rede regular com grau $k = 2$ e $n = 10$ e sua distribuição de grau.

REDES ALEATÓRIAS Redes aleatórias, muito estudadas na matemática, são redes cujas arestas são criadas por um processo aleatório. Comumente, uma rede aleatória mantém alguns parâmetros fixos, como número de nós, porém com uma construção estocástica. Entre os modelos de redes aleatórias mais estudados encontra-se o modelo de Gilbert, denotado por $G(n, p)$ onde n é o numero de nós e toda possível aresta entre esses n nós é criada independentemente com uma probabilidade $0 < p < 1$[20]. Outro modelo muito estudado foi proposto por Paul Erdös e Alfréd Rényi, o modelo Erdös-Rényi, onde uma rede aleatória é denotada por $G(n, m)$ onde n é o número de



(a) Rede Regular (b) Distribuição de Grau

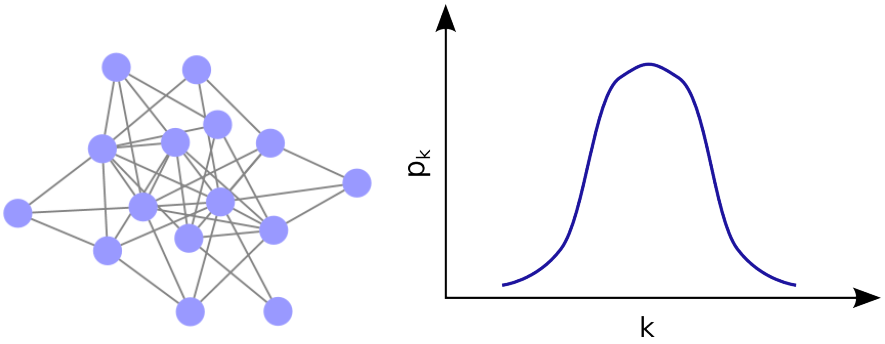
Figura 5: Rede regular com $k = 2$, $n = 10$ e $m = 10$.

nós e $0 < m < \frac{n(n-1)}{2}$ o número de arestas em uma rede simples[6]. A distribuição de grau dos nós de uma rede aleatória segue uma distribuição binomial

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{4}$$

centrada no grau médio da rede $\langle k \rangle = (n-1)p$ para um modelo de Gilbert e centrada no grau médio $\langle k \rangle = \frac{2m}{n}$ para um modelo Erdös-Rényi. Para $n \to \infty$, a distribuição de grau pode ser denotada por uma distribuição de Poisson.

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{5}$$



(a) Rede Aleatória (b) Distribuição de Grau

Figura 6: Rede aleatória com $n = 15$ e $m = 40$. O tamanho do vértice é proporcional ao seu grau.

REDES LIVRE-DE-ESCALA Muitos sistemas, naturais ou artificiais, quando modelados em forma de rede, apresentam uma topologia caracterizada por conter poucos nós com alto grau, chamados *hubs*, e muitos nós com baixo grau chamada livre-de-escala. A nível de exemplo, diversas diversos sistemas assim modelados como redes sociais,

redes de computadores, a internet, rede de aeroportos e redes metabólicas apresentam esta topologia. O surgimento dessa topologia em tantos sistemas foi elucidada em um modelo de construção dessas redes proposto por Albert-Lázló Barabási e Réka Albert onde a rede é criada por um comportamento de "*the rich get richer*[3]", ou seja, nós com maior grau tem maior preferência a receber novos links. No processo de construção de uma rede G com $n > 2$ nós, se inicia o processo com dois nós conectados e, a cada nó adicionado na rede, há uma probabilidade

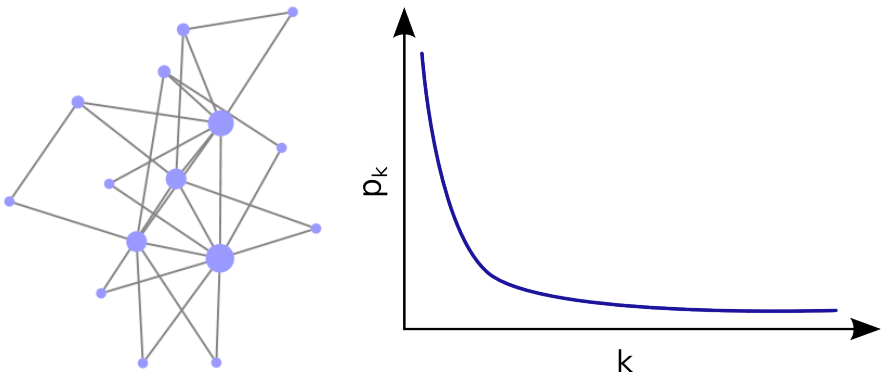
$$p_i = \frac{k_i}{\sum_0^j k_j} \tag{6}$$

de se conectar ao nó i com j sendo o número de nós presentes na rede.

A distribuição de grau dos nós de uma rede com topologia livre-de-escala, com $n \to \infty$, segue uma lei de potência

$$p_k \sim k^{-\alpha} \tag{7}$$

com o parâmetro α estando comumente no intervalo $2 < \alpha < 3$ na maioria dos sistemas mas podendo também ocorrer valores diferentes.



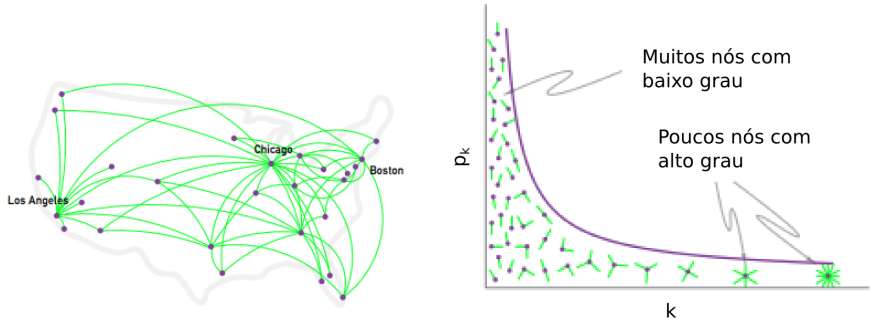
(a) Rede Livre-de-Escala (b) Distribuição de Grau

Figura 7: Rede livre-de-escala com $n = 15$ e $m = 27$. O tamanho do vértice é proporcional ao seu grau.

A topologia livre-de-escala revela algumas propriedades emergentes do sistema como robustez à falhas aleatórias e ao mesmo tempo, uma fraqueza contra falhas direcionadas. Pode-se tomar como exemplo um sistema simplificado da rede de aeroportos dos Estados Unidos. Se um aeroporto for escolhido aleatóriamente com a mesma probabilidade para todos e for fechado, a chance de prejudicar o sistema como um todo é baixa, pois a quantidade de aeroportos com poucas conexões é muito maior do que a quantidade de aeroportos principais. Por outro lado, se for fechado intencionalmente um certo aeroporto,

3 Tradução: os ricos ficam mais ricos.

como o de Chicago por exemplo, a estrutura do sistema seria seriamente abalada mesmo com a remoção de um único nó que tem o papel de *hub*, ou seja, tem muitas conexões.



(a) Rede de Aeroportos (b) Distribuição de Grau

Figura 8: Rede de aeroportos dos Estados Unidos simplificada. Apresenta uma topologia livre-de-escala podendo observar a característica de robustez à falhas aleatórias com uma quantidade muito maior de nós com baixo grau. Ao mesmo tempo, é uma estrutura com pontos fracos se um ataque direcionado for disparado contra seus hubs. Figura obtida do livro "*Network Science*"de Albert-Lazlo Barabási.

Caracterização de Redes

No estudo de redes complexas, diversas medidas aplicadas às redes são utilizadas para extrair informações sobre diferentes aspectos do sistema. Além do grau dos nós denotado por k já mencionado anteriormente, medidas como diâmetro, número de componentes, mínimo caminho, caminho mínimo médio, *betweenness* e coeficiente de clusterização são medidas básicas necessárias para caracterizar e entender a estrutura de um sistema. Nesta seção serão abordadas algumas das medidas mais utilizadas.

CAMINHO MÍNIMO MÉDIO E DIÂMETRO Dado dois nós em uma rede, o mínimo caminho, denotado por l, é o caminho de distância ℓ mínimo os liga, em outras palavras, é o número mínimo de arestas percorridas pela rede para uní-los, também chamado de caminho geodésico. No caso de dois nós estarem em componentes diferentes da rede, ou seja, não há nenhum caminho que os liguem, o valor de l é convencionalmente definido como 0. A média dos mínimos caminhos entre todo par de nós de uma rede é calculado por

$$\langle l \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} l_{ij} \forall l_{ij} > 0 \tag{8}$$

com $i \in n$ e $j \in n$ podendo ser aproximada para

$$\langle l \rangle = \frac{1}{n^2} \sum_{i \neq j} l_{ij} \tag{9}$$

no caso do limite de um n grande. A utilização mais famosa desta medida foi feita por Jeffrey Travers e Stanley Milgram em um experimento de sociologia onde queriam medir o caminho mínimo médio entre duas pessoas nos Estados Unidos por intermédio de uma rede social. O valor encontrado foi de 6, cunhando a famosa expressão "seis graus de separação"[21]. Estes estudos se baseiam no conceito de "*small-world networks*[4]", característica encontrada em diversas redes de sistemas reais onde mesmo havendo um número muito elevado de nós, a distância entre eles costuma ser pequena. No ano de 2011, a empresa Facebook divulgou o caminho mínimo médio medido entre 721 milhões de usuários por ligações de amizade na rede social encontrando um valor de 4.74. Assim como a média, um histograma dos mínimos caminhos pode também trazer informações interessantes.

Outra utilização do mínimo caminho l é para o cálculo do diâmetro da rede, o qual consiste no maior dos mínimos caminhos entre os nós de uma rede. O diâmetro de uma rede é muito utilizado para analisar o comportamento da estrutura de uma rede durante alterações ao longo do tempo em uma dinâmica.

BETWEENNESS Um *hub* dentro de uma rede certamente tem sua importância tanto na estrutura da rede quanto na dinâmica do sistema modelado. Contudo, o número de conexões nem sempre revela todos os nós ou elementos importantes de uma rede ou sistema. Suponha-se que existem duas grandes rodoviárias, A e B em duas distantes e importantes cidades, as quais fazem conexões com muitos outros lugares. Suponha que para chegar de uma rodoviária a outra, de A a B, é necessário pegar dois ônibus os quais fazem conexão em uma pequena rodoviária C de uma pequena cidade no caminho. Mesmo C sendo uma pequena rodoviária a qual faz conexão com apenas duas cidades, sem ela não é possível transitar entre as duas grandes rodoviárias A e B. Neste sentido, a medida de *betweenness* mede a importância de um nó (ou também de uma aresta) na conectividade de uma rede e também seu poder no controle de uma informação que transita pela por ela. Matematicamente, podemos expressar o *betweenness* b de um nó i pela expressão

$$b_i = \frac{1}{n^2} \sum_{s \neq t} q^i_{st} \tag{10}$$

onde q^i_{st} é 1 se o caminho geodésico entre os nós s e t passa pelo nó i e 0 se o caminho não passa pelo nó i ou não existe caminho entre os

4 Tradução: redes mundo pequeno.

nós s e t. O valor da soma dos caminhos que passam por i é normalizado por $\frac{1}{n^2}$ para um n suficientemente grande mantendo $0 \leqslant b_i \leqslant 1$. Entretanto, uma rede pode conter mais de um caminho geodésico entre dois vértices, ou seja, podem existir mais de um caminho com a mesma distância que ligam dois nós. Neste caso, podemos adotar um peso $\frac{1}{g_{st}}$ para cada caminho geodésico existente entre s e t onde g_{st} é o número de caminhos geodésicos existentes. Desta forma, a equação se modifica para

$$b_i = \frac{1}{n^2} \sum_{s \neq t} \frac{g_{st}^i}{g_{st}} \tag{11}$$

COEFICIENTE DE CLUSTERIZAÇÃO Enquanto o *betweenness* trata de quantificar a importância de um nó na conectividade ou no controle de informações que transitam numa rede em um contexto amplo, o coeficiente de clusterização objetiva medir estas capacidades em um contexto mais reduzido, limitando-se aos vizinhos de cada nó. Esta medida c de um nó i pode ser obtida por

$$c_i = \frac{C_i}{\frac{1}{2} k_i (k_i - 1)} \tag{12}$$

onde C_i é o número de pares de vizinhos de i que são conectados entre si. Também podemos escrever esta equação em termos de uma medida de redundância proposta por Burt[22] e simplificada por Borgatti tendo que a redundância R de um vértice i é dada por

$$R_i = \frac{1}{k_i} \sum_j k_{ji} \tag{13}$$

onde k_{ji} é o número de conexões que um vizinho j de i faz com outros vizinhos de i[23]. Desta forma, podemos reescrever C_i em termos de R_i sendo $C_i = \frac{1}{2} k_i R_i$ e então temos

$$c_i = \frac{\frac{1}{2} k_i R_i}{\frac{1}{2} k_i (k_i - 1)} = \frac{R_i}{k_i - 1}. \tag{14}$$

Como $k_i - 1$ é o maior valor possível de R_i, c_i encontra-se entre 0 e 1. No contexto de redundância em uma rede, como desejado por exemplo em vias terrestres para disponibilizar diferentes rotas no caso de sobrecarga de uma delas, quanto mais próximo de 1 o valor de c_i, melhor seria o fluxo de tráfego para esta região. Em um contexto de controle de informação, um valor próximo de 0 de c_i aumenta o poder de controle de passagem de informação por aquele nó.

COMUNIDADES Entidades muito relacionadas dentro de um sistema muitas vezes compartilham características, objetivos ou funções formando grupos ou comunidades. Na sociedade, por exemplo, pessoas formam grupos por amizades, interesses, religião entre outros

motivos. Identificar comunidades nas redes que modelam estes sistemas se mostra uma ferramenta muito útil no seu entendimento e manipulação e é um grande e muito ativo campo de pesquisa. A pesar de não existir uma única definição para uma comunidade e podendo ser abstraída em diversos níveis, podemos considerar como uma comunidade um grupo de vértices dentro de uma rede tal qual possua mais conexões entre seus membros internos do que conexões com os membros externos ao grupo[24]. Existem atualmente diversas técnicas para detecção de comunidades em redes com diferentes abordagens e níveis de complexidade baseadas em medidas de modularidade, heurísticas e particionamento de grafos dentre outras.

Considerações Finais

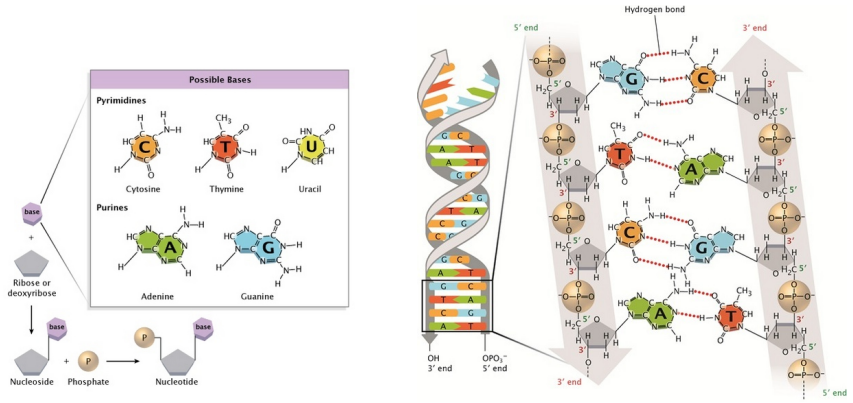
Nesta seção foram abordados tópicos básicos dentro da teoria de redes complexas os quais servirão de base computacional e matemática para o desenvolvimento do projeto proposto, uma vez que o sistema a ser analisado será modelado em formato de rede.

As medidas aqui estudadas serão aplicadas à rede de uma célula completa gerada no presente projeto com o objetivo de caracterizá-la matematicamente e extrair informações biológicas relevantes a partir de medidas matemáticas.

BIOLOGIA MOLECULAR

Dogma Central da Biologia

Toda célula tem a capacidade de transmitir suas características genéticas a suas células filhas concedendo-as um material genético idêntico ou quase idêntico da célula mãe. Esta informação genética se encontra codificada em forma de uma molécula de DNA (ácido desoxiribonucléico) a qual é uma macromolécula polimérica composta de pequenos blocos chamados nucleotídeos, compostos por um açúcar, um grupo fosfato e uma base nitrogenada, sendo eles basicamente quatro: adenina (A), timina (T), citosina (C) e guanina (G) ilustrados pela Fig. 9a. O DNA se encontra no formato de uma dupla hélice formada por sequências desses nucleotídeos ligados por ligações fosfodiéster em comprimento e as duas fitas de sua dupla hélice são ligadas por pontes de hidrogênio por complementariedade de bases nitrogenadas sendo A complementar a T e C complementar a G[25] como ilustrado na Fig. 9b. Na sequência de nucleotídeos em cada fita está codificado o código genético a ser transmitido para as células filhas no processo de duplicação celular. No processo de divisão celular, o DNA é replicado em duas cópias idealmente idênticas garantindo que cada célula filha tenha uma cópia íntegra de todo o material genético.



(a) Estrutura química dos nucleotídeos (b) Estrutura da dupla hélice do DNA

Figura 9: Os nucleotídeos adenosina, timina, citosina e guanina se combinam em sequências não uniformes formando duas hélices complementares, as quais formam uma molécula de DNA. Figuras retiradas de *Nature Education*[26].

A informação genética codificada no DNA contém toda a informação necessária para a produção de proteínas, moléculas que compõem a estrutura e maquinaria molecular das células, entretanto, apesar de ser uma forma estável, ainda é uma forma estática de armazenamento de informações. Para que essa informação seja de fato executada, sua sequência é transcrita em moléculas chamadas RNAs (ácido ribonucléicos), compostos pelos mesmos nucleotídeos que o DNA havendo três diferenças principais em relação ao DNA: todos os nucleotídeos possuem um grupo hidroxila a mais em seu açúcar, o nucleotídeo timina (T) é substituído por um diferente nucleotídeo chamado uracila (U) e a molécula é composta por apenas uma fita. Essas moléculas levam para outros locais da célula a informação necessária copiada do DNA para que sejam sintetizadas as proteínas.

Uma vez sintetizados e processados, os RNAs se encontram com uma das moléculas mais abundantes dentro de uma célula, o ribosomo. O ribossomo é uma riboproteína, ou seja, é composta por aminoácidos e RNA e tem como função sintetizar novas proteínas a partir do código obtido em um mRNA (RNA mensageiro). Este código podemos didaticamente exemplificar como uma sequência linear de letras (A,U,C e G) e é lido pelo ribossomo em conjuntos de três letras, chamados códons. Cada códon codifica um aminoácido durante a síntese da proteína onde o ribossomo catalisa a ligação desse novo aminoácido na proteína a ser formada e então passa para o próximo códon. Ao fim desse processo, uma nova proteína é liberada, esta podendo passar por modificações extras ou estar pronta para exercer sua função na célula. As proteínas são os principais constituintes de uma célula podendo exercer as mais diversas funções como estrutura, sinalização, catálise de reações, sínteses poliméricas entre diversas outras.

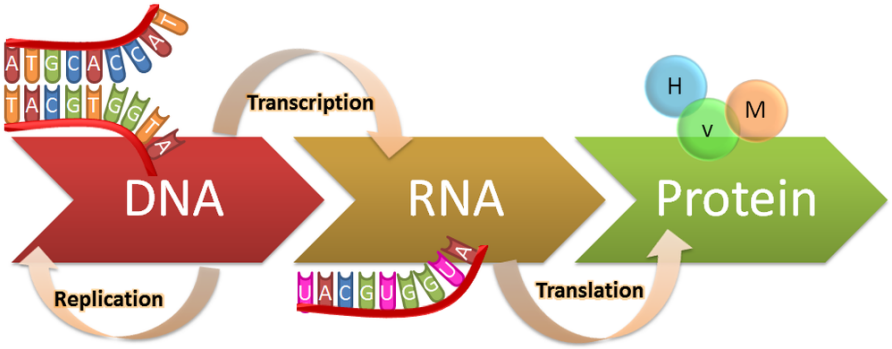


Figura 10: Fluxo de replicação e transmição da informação genética armazenada no DNA, sendo transmitida por meio de moléculas de RNA para então ser traduzida em proteínas, moléculas funcionais da célula. Figura obtida em [**?**]

Em resumo, a capacidade de replicação e a transmissão de informação codificada no DNA, transcrita para RNA e então traduzida para proteína são conhecidas como o dogma central da Biologia, termo criado em 1970 por um dos descobridores da estrutura do DNA, Francis Crick[27]. Este processo é tido como um dogma central pois ocorre em toda célula desde simples bactérias até as complexas células do corpo humano assim constituindo a base de todo organismo vivo.

Subsistemas Celulares

Além dos processos de replicação, transcrição e tradução contidos no dogma central da Biologia, descritos na seção 2.2.1 desse documento, outros processos igualmente essenciais co-ocorrem nas células como absorção e processamento de nutrientes, captação e emissão de sinais externos, reparo de danos, entre outros, para garantir que a célula se mantenha viva e tenha condições para se replicar. A seguir, alguns dos mais importantes processos serão listados e brevemente descritos para melhor contextualizar um ambiente intracelular.

METABOLISMO Para que a célula possa crescer, se replicar ou mesmo somente manter em perfeito funcionamento seus outros processos, é necessário que consuma nutrientes no meio em que está e os processe para produzir energia (majoritariamente em forma de moléculas de ATP) e outros metabólitos a serem utilizados por outros processos. O processo de absorção e transformação de nutrientes é chamado de metabolismo e é constituído de um conjunto de reações bioquímicas catalisadas por proteínas chamadas de enzimas. Uma das vias metabólicas mais conhecidas é a da glicólise onde as moléculas de glicose absorvidas pela célula passam por uma série de transformações cata-

lisadas por enzimas a fim de produzir duas moléculas de ATP para cada molécula de glicose.

SINALIZAÇÃO Para responder a estímulos e alterações do ambiente, certas proteínas tem a capacidade de reconhecer esses estímulos e enviar informações para o interior da célula acarretando em mudanças na expressão gênica ou no funcionamento de outros processos. Estes sinais são normalmente captados por proteínas inseridas na membrana chamadas de receptores os quais geram uma modificação, como uma fosforilação, em uma proteína complementar na parte interna da célula. Essa complementar interna da início a uma cascata de modificações em outras proteínas, como por exemplo uma cascata de fosforilação por kinases, onde essas modificações podem incluir fatores de transcrição ou enzimas.

TRANSPORTE A membrana celular limita o espaço inerente à célula criando um meio interno, ou intracelular, e um meio externo, ou extracelular. Algumas células também possuem divisões de compartimentos internos por membranas. A célula durante sua vida necessita que nutrientes e outras moléculas sejam absorvidas para o meio intracelular, proteínas e subprodutos do metabolismo sejam transportados para o meio extracelular, como também o transporte de proteínas e outras moléculas entre seus compartimentos internos. Estes processos, chamados de transporte podem se dar em duas diferentes formas: transporte ativo, quanto há gasto de energia; e transporte passivo, quando não envolve gasto de energia.

PROCESSAMENTO DE RNAS Após o processo de transcrição, alguns RNAs necessitam passar por um processamento. Em organismos eucarióticos, é comum um gene ser composto por introns e exons. Os introns são trechos de DNA que não codificam proteínas e necessitam ser removidos da sequência de RNA transcrita em um processo chamado *splicing* envolvendo enzimas chamadas endonucleases e ligases que tem como função cortar e ligar fitas de RNA respectivamente. Em outros casos, uma fita de RNA transcrita pode conter informação de mais de uma proteína ou conter trechos com diferentes funções no caso de tRNAs (RNAs transportadores) e necessitando ser seccionada em seus trechos funcionais.

MODIFICAÇÃO DE PROTEÍNAS Algumas proteínas necessitam sofrer modificações após seu processo de síntese para que possam desempenhar sua função corretamente. Essas modificações podem envolver adição de grupo prostéticos (a exemplo da hemoglobina), adição de açúcares, adição de lipídeos (como em proteínas de membrana) fosforilações de aminoácidos específicos entre outras.

DEGRADAÇÃO Com fins de regulação de atividade celular e reciclagem de nutrientes, RNAs e proteínas podem ser degradados em seus blocos de construção: nucleotídeos e aminoácidos respectivamente. Este processo envolve enzimas chamadas proteases e peptdases, no caso de degradação de proteínas, e RNAses e nucleases no caso de degradação de RNAs. A degradação pode ocorrer em diferentes cenários como exemplo a redução do excesso de alguma molécula regulando sua atividade, obtenção de nutrientes para construção de novas moléculas ou simplesmente degradação de moléculas velhas.

REPARO DE DNA A molécula de DNA, apesar de muito estável podendo manter sua estrutura por milhões de anos, é passível de ser danificada por agentes externos. Esses agentes envolvem calor excessivo, radiação, moléculas de impregnação intra-DNA e também no próprio processo de replicação podem ocorrer erros sendo inseridos nucleotídeos errados na sequência. Para contornar esse problema, muitas células se dispoem de proteínas muito especializadas as quais conseguem identificar erros na estrutura do DNA e corrigí-los.

Considerações Finais

Neste capítulo foi abordado o objeto de estudo deste projeto, a célula, e os principais sistemas que a compõem. Tendo como base a biologia molecular, os tópicos abordados neste capítulo introduzem de forma geral o objeto a ser modelado durante e descrevem brevemente sua importância e principais características. Os processos aqui descritos estão presentes em toda célula, não sendo específico para uma célula em particular.

Neste projeto propõe-se a modelagem de todos os sistemas que compõem uma célula, dado um organismo específico descrito posteriormente neste documento, de forma a construir um modelo integrativo desses sistemas o qual possa representar a célula como um todo.

REDES BIOLÓGICAS

O grande volume de dados gerado pela intensa pesquisa na área de biologia molecular proporciona material para poder dar mais um passo no entendimento de organismos vivos, possibilitando abordagens sistêmicas envolvendo mais do que poucas moléculas como se tem feito até então. Desta forma, a modelagem de sistemas biológicos em formato de redes vem sedo utilizada para elucidar padrões de interação e fluxos de informações em diversos campos da biologia molecular. Neste capítulo serão abordados os modelos de redes biológicas utilizados na biologia molecular, métodos que geram dados para esses modelos e algumas das aplicações mais importantes.

Redes Metabólicas

O metabolismo de uma célula é responsável pela criação e degradação de moléculas (metabólitos) que servem como nutrientes para outros processos celulares como citado na seção 2.2.2. Sua melhor representação em formato de redes é composta por um conjunto de nós representando metabólitos e outro conjunto de nós representando as reações bioquímicas. Estes grupos de nós são ligados entre si por links direcionados indicando quais metabólitos são reagentes (links que apontam para nós reações) e quais são produtos (links que apontam para metabólitos) de uma reação. Portanto, trata-se de um modelo bipartido e direcionado de rede. Em alguns modelos, as enzimas que catalisam estas reações são ligadas por um link não direcionado aos nós que representam reações catalisadas por ela, uma vez que não são produzidas nem consumidas. Desta forma, as reações bioquímicas são ligadas em cadeias formando *pathways* ou vias metabólicas. Outra forma de representação, a qual mesmo perdendo informação é muito utilizada, se obtém com uma rede simples direcionada transformando a rede bipartida em sua projeção de um modo somente com o grupo de nós que representam metabólitos.

Os dados necessários para a construção de redes metabólicas provém de demorados e onerosos ensaios bioquímicos caracterizando a atividade de cada enzima presente nas vias e estimando seus reagentes e produtos. Entretanto, muitos esforços já foram realizados para gerar tais dados sobre uma vasta gama de organismos. Redes metabólicas podem ser encontradas em bancos de dados gratuitos como MetaCyc e BioCyc[28] contendo 2411 vias metabólicas até 2014. A Fig. 11 demonstra a rede metabólica do organismo *Escherichia coli* obtido no website do BioCyc. Outra fonte onde se pode encontrar redes metabólicas é o KEGG (Kyoto Encyclopedia of Genes and Genomes)[29], porém com conteúdo privado disponível sob pagamento.

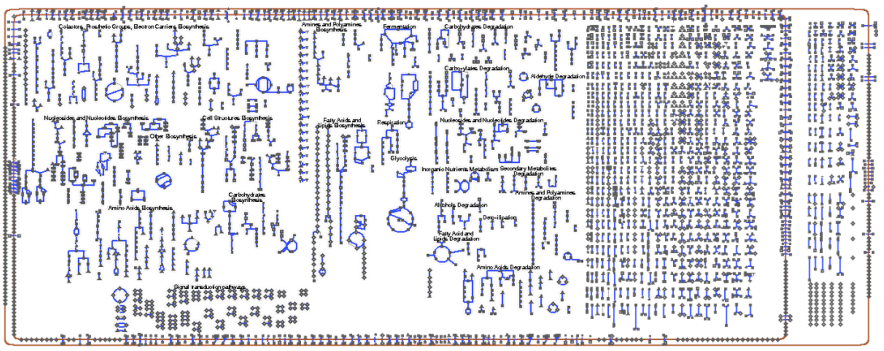


Figura 11: Rede metabólica disponível no website do BioCyc contendo todas as reações bioquímicas conhecidas no metabolismo do organismo *Escherichia coli*.

Dentre as diversas aplicações de redes metabólicas, as que se destacam são na otimização de processos bioquímicos em indústrias em geral e principalmente indústrias farmacêuticas[30]. Também serve como base para métodos de simulação como FBA (*Flux Balance Analysis*) onde se pode prever o fluxo de produção de metabólitos em uma célula tanto num contexto estático quanto dinâmico[31].

Redes de Interação Proteína-Proteína

Proteínas se relacionam de diversas formas, dentro e fora de uma célula, podendo formar complexos proteicos, catalisar modificações em outras proteínas, transportar proteínas, agir em conjunto em uma via metabólica e etc. As relações proteicas onde duas ou mais proteínas se ligam fisicamente para desenvolver alguma atividade podem ser utilizadas para construir redes chamadas "redes de interação proteína-proteína" ou "PPI networks"[5] onde os nós são compostos por proteínas e podem ser ligados entre si caso haja alguma interação entre elas. Os links podem ter pesos de acordo com a quantidade de evidências daquela interação entre as proteínas podendo delimitar um grau de confiabilidade para a rede. Portanto, interações entre proteínas são normalmente representadas por uma rede ponderada e não direcionada podendo também haver representações com links direcionados indicando tipos de interações.

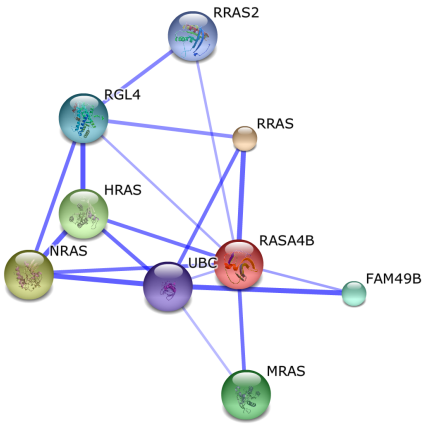


Figura 12: Rede de interação proteína-proteína da RASA4B humana(*Homo sapiens*), uma GTPase de baixo peso envolvida na via de sinalização Ras-MAPK.

Para identificar estas interações são utilizados métodos como "*two-hybrid screening*", *microarrays* de proteína, co-imunoprescipitação, purificação por afinidade com espectrometria de massas, dentre outras. Contudo, todas essas técnicas correm um alto risco de gerarem falsos negativos e falsos positivos. Redes PPI podem ser encontradas no

5 Protein-Protein Interaction Networks

banco de dados STRING[32] encontrando uma boa interface em seu website para analisá-las e também no banco de dados do BioGRID[33]. A Fig 12 mostra um exemplo de rede PPI obtida no STRING-db.

A análise de redes PPI tem grande importância na prospecção de alvos para drogas no tratamentos de diversas doenças, incluindo o câncer[34]. Proteínas com papéis centrais na célula podem acarretar diversos efeitos negativos caso não funcionem corretamente como o caso da mutação no gene p53, muito correlacionado com crescimento de tumores[35]. Por outro lado, proteínas que desenvolvem um importante papel dentro de microrganismos patológicos podem oferecer importantes alvos para desenvolvimento de antibióticos.

Redes Gênicas

A maquinaria molecular de que a célula dispõe para cumprir sua função e se multiplicar é criada com base em seu genoma como discutido na seção 2.2.1. Contudo, nem todos os genes são necessariamente expressos a todo momento, a célula produz suas proteínas de acordo com a demanda ou fase do ciclo celular. Essa regulagem da expressão gênica é feita através de proteínas chamadas de "fatores de transcrição" os quais se ligam em regiões promotoras de genes específicos podendo incentivar ou inibir a expressão do gene. Uma vez que essas proteínas são produtos de outros genes, pode se indicar a relação entre esses genes no âmbito da regulagem de suas expressões. Este sistema de regulagem da expressão gênica é comumente modelado em formato de rede onde os vértices representam genes e os links entre os vértices podem representar dois tipos de interação: aumentar ou inibir a expressão gênica. Desta forma, redes de regulação de expressão gênica são representadas por redes direcionadas com dois tipos de aresta.

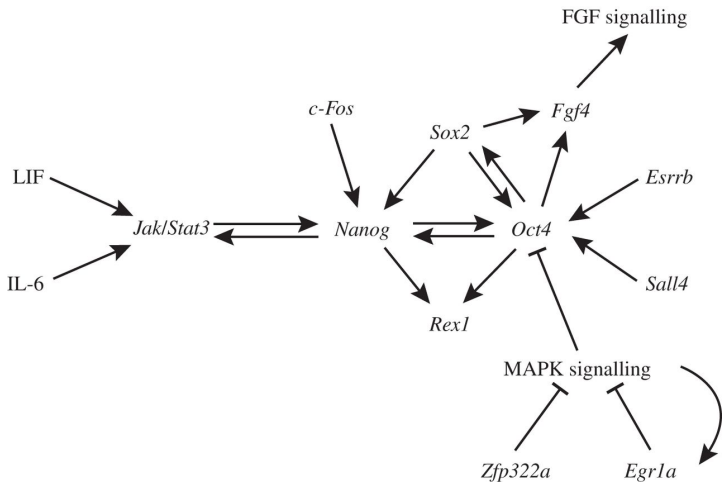


Figura 13: Rede de regulação gênica de células tronco no desenvolvimento embrionário de ratos (*Mus musculus*)[36].

O processo para identificar a estrutura de uma rede gênica envolve diversos ensaios como eletroforese, DNA *footprint*, *microarrays* e também técnicas computacionais de bioinformática para analisar possíveis sequências promissoras à receber fatores de transcrição.[37] A confiabilidade desses processos não permite determinar com exatidão essas interações, portanto, gerar dados para alimentar esse tipo de rede é custoso e demorado. Os dados gerados até então para diversos organismos e linhagens celulares podem ser encontrados em banco de dados como GeneNetwork, EsyN, HumanMine, KEEG, Reactome.org, sendo estes os mais conhecidos.

Os dados fornecidos por uma rede de regulação de expressão gênica juntamente com dados sobre níveis de expressão servem como base para simulações computacionais de expressão gênica podendo prever em alguns casos comportamento celular[38]. Estas redes também são de grande valia no entendimento do funcionamento de um sistema vivo mas ainda são necessários muitos avanços para que se possam obter mapas mais confiáveis e de maior porte.

Interactomas

Recentes avanços em equipamentos de *high-throughput* geram uma grande quantidade de dados os quais são utilizados para construir redes biológicas. Quando redes tomam escala de representar sistemas de escala celular, não somente de certos processos particulares, são comumente chamados de interactomas. Este termo, apesar de poder ser utilizado para quaisquer rede biológica-molecular de grande porte como redes metabólicas de células completas como na Fig. 11, redes gênicas de escala celular, porém, se refere mais comumente a redes PPI de larga escala como ilustrado na Fig 14 o interactoma de proteínas humanas.

Interactomas podem ser encontrados em diversos bancos de dados como Interactome.org, BioGRID, EBI IntAct e Reactome.org. Mesmo que os interactomas representem informações à escala celular, ainda modelam a célula de diferentes perspectivas, não contemplando o todo, mesmo que todos os interactomas de um organismo estejam de fato interligados. Esta não integração de interactomas deixa espaço aberto para novas pesquisas no quesito de integração de dados.

Considerações Finais

Nesta seção foram abordados os principais tipos de redes biológicas utilizados atualmente assim como seus métodos de construção, representação e aplicações. Todos estes modelos estão em constante pesquisa e geração de dados tendo-se mostrado um campo muito frutífero. A análise desses modelos biológicos também revelam as barreiras e dificuldades a serem superadas nesta área e um promis-

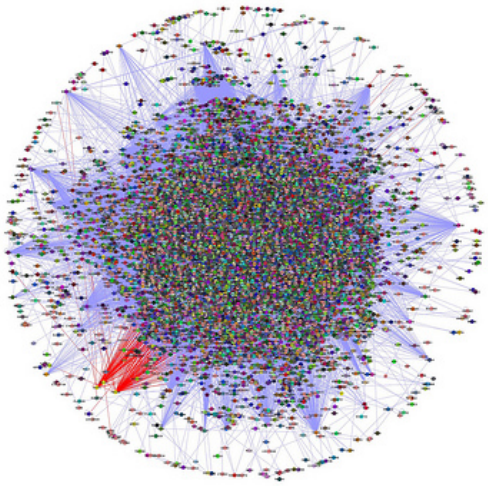


Figura 14: Interactoma de interações proteicas humanas. Créditos da imagem à Andrew Garrow.

sor futuro no entendimento holístico do comportamento e estrutura de sistemas vivos.

Parte II

DESENVOLVIMENTO DO PROJETO

3

RESULTADOS

Neste capítulo serão apresentados os resultados obtidos na formulação e análise de um modelo computacional de sistemas celulares completos.

MODELAGEM INTEGRATIVA DE SUBSISTEMAS CELULARES

As células, apesar de serem comumente estudadas sendo divididas por grupos funcionais de moléculas ou subsistemas celulares, na prática todos esses subsistemas estão conectados entre si e funcionam de forma conjunta e harmônica em um quadro saudável. Como embasado nas seções 2.2 e 2.3, esses subsistemas são modelados de forma reducionista em diferentes tipos de rede tornando a informação, ainda que sobre o mesmo organismo, heterogênea, dificultando a busca por informações integrativas e abordagens mais holísticas para um melhor entendimento da dinâmica celular.

Com o objetivo de obter uma modelagem integrativa e homogênea que tenha a capacidade de englobar em um único modelo todo tipo de subsistema celular, foi definido um conjunto de regras a serem seguidas na construção de tal modelo em formato de rede, como embasado na seção 2.1. Uma rede gerada segundo estas regras chamamos de *Whole-Cell Network* (Rede de Célula Completa) e possui dois tipos de nós: nós molécula e nós reação. As regras a serem seguidas na construção da rede são:

- Cada molécula ou estrutura em uma célula deve ser representado por um único nó molécula;
- Cada estado diferente de uma molécula (ex., uma proteína ativa ou inativa) deve ser representado por um nó molécula diferente;
- Cada interação molecular deve ser representada por um único nó reação;
- Reações com múltiplos passos (ex., polimerização, degradação) devem ser condensadas em um único nó reação;
- Conexões devem ocorrer somente entre tipos diferentes de nós;
- Conexões que apontam nós reação podem ser de dois tipos: conexão reagente, o qual liga nós molécula que serão consumidos durante as reações, e conexão modificadora, a qual liga nós molécula que representam moléculas catalíticas ou que não se modificam durante as reações;

- Conexões que partem de nós reação ligam a moléculas produzidas pelas reações;
- O peso de cada conexão determina a estequiometria da interação.

Por definição, uma rede gerada seguindo essas regras será bipartida, direcionada e ponderada. Estas regras permitem a modelagem de qualquer interação molecular a nível celular desde uma simples reação bioquímica até o complexo processo de divisão celular. Para ilustrar a modelagem de interações moleculares a Fig. 15I exemplifica a modelagem de (a) uma reação bioquímica $Met_1 + Met_2 \xrightarrow{Enz} Met_3$; (b) inibição de uma proteína por um ligante; (c) uma reação de polimerização, como a síntese de uma proteína catalisada por uma enzima; (d) transporte de moléculas do compartimento 1 para o compartimento 2 por um agente transportador; (e) formação de um complexo proteico composto de duas proteínas.

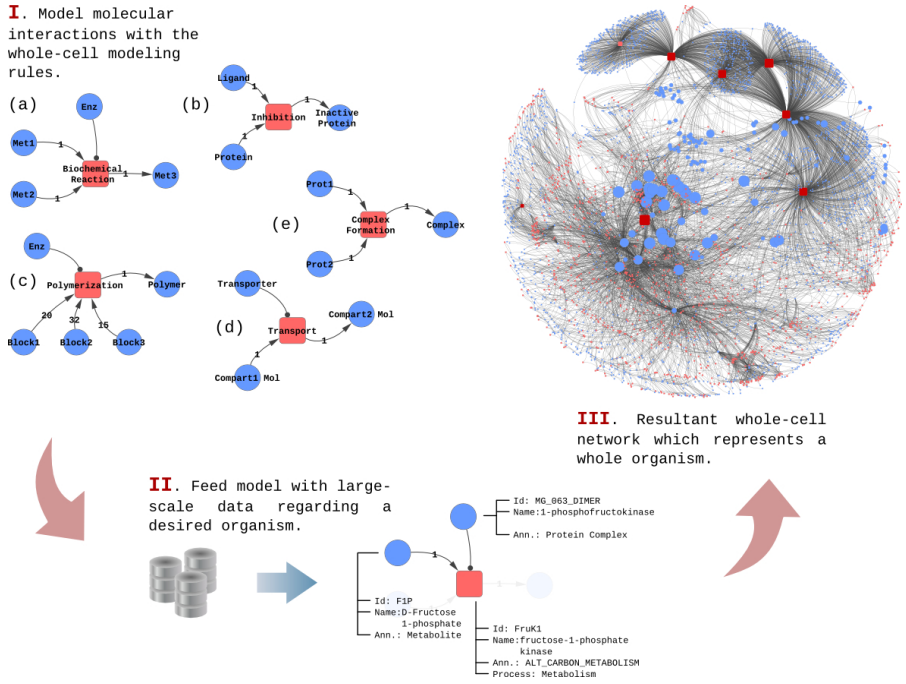


Figura 15: Processo de construção de uma *Whole-Cell Network*: (I) a construção se inicia modelando os processos celulares respeitando as regras definidas nesta seção; (II) A rede é então alimentada com informações de larga escala sobre o organismo desejado. Este processo cria nós molécula para todas as moléculas e estruturas da célula assim como nós reação para cada interação conhecida; (III) A rede resultante representa todas as interações conhecidas entre moléculas e estruturas de uma célula completa integrando todos os processos celulares. Os círculos azuis representam nós molécula e os quadrados vermelhos representam nós reação.

ESTUDO DE CASO

Para um estudo de caso, foi escolhido como modelo biológico a bactéria *Mycoplasma genitalium*, uma bactéria Gram-positiva, patogênica e detentora do menor genoma conhecido até a presente data contendo 580kb em extensão. Esta escolha foi baseada em (1) sua relativa simplicidade, contendo somente 525 genes em comparação com aproximadamente 30.000 genes de uma célula humana, (2) por sua importância médica, sendo causadora de infecções na região urogenital[39] e (3) pela disponibilidade de estudos acerca de sua dinâmica celular[40, 18] e anotação de seu genoma incluindo um banco de dados com informações curadas sobre o organismo e organismos homólogos[41].

Aquisição de dados

Para a construção da *Whole-Cell Network* do organismo *M. genitalium* foi utilizado o banco de dados WholeCellKB[41] em MySQL, o qual contém informações sobre todos os processos celulares, moléculas e estruturas conhecidos sobre a bactéria *M. genitalium*. Este banco de dados contém informações obtidas em mais de 900 artigos científicos, sendo assim, a confiabilidade das interações modeladas depende da confiabilidade do banco de dados. A fim de definir as estequiometrias das reações de transcrição, foi utilizado o genoma da cepa *Mycoplasma genitalium* G37 obtido no banco de dados do NCBI, referência `NC_000908.2`. As estequiometrias das reações de tradução foram obtidas a partir das sequências de aminoácidos traduzidos das sequências gênicas do genoma supracitado utilizando software EMBOSS Transeq[42] rodando localmente. As informações contidas no banco de dados rodando localmente, como unidades transcricionais, genes, proteinas, complexos proteicos, metabólitos e reações foram acessadas utilizando um script em Python 2.7. Utilizando a API para Python LibSBML, foi criado um modelo em SBML onde todas as moléculas e estruturas encontradas no banco de dados foram armazenadas como objetos `Species` e todas as reações encontradas foram armazenadas como objetos `Reaction` conectando em si os objetos `Species` correspondentes. Enzimas e moléculas catalíticas foram ligadas às reações como objetos `Modifier`[1].

Modelagem e Construção da Rede

Uma vez extraídos os dados do banco de dados como descrito na seção 3.2.1, foi necessário realizar uma manipulação dos mesmos para entrarem em conformidade com as regras de modelagem de uma *whole-cell network* definidas na seção 3.1.

1 Essas moléculas serão referidas como "modificadores" nas próximas seções.

Alguns dados obtidos do banco de dados, como por exemplo as reações bioquímicas do metabolismo, já estavam em conformidade pois ja são comumente estruturados em forma similar a modelagem feita neste trabalho, como exemplificado na seção 2.3.1. Para outros dados, sua modelagem respeitando as regras definidas neste trabalho não eram tão intuitivas. Para isso, os próximos parágrafos descrevem a modelagem de estruturas ou processos particulares do organismo *M. genitalium*.

REPRESENTAÇÃO DO CROMOSSOMO Cada unidade transcricional, como um gene ou um polycistron, foi representado por um nó molécula. A replicação do cromossomo se inicia com o OriC formando um complexo com o polímero de DnaA para assim formar o nó "*replication initiation complex*[2]". A reação de polimerização da nova fita de DNA tem como produto um nó cromossomo. Esse nó cromossomo participa sequencialmente das reações de dano ao DNA, reparo de DNA, segregação e e finalização da replicação.

REPRESENTAÇÃO DA MEMBRANA A membrana celular foi representada como um único nó molécula o qual funciona como um modificador em toda reação que acontece na membrana. Por exemplo, o nó membrana é um modificador em todas as reações de transporte na membrana, reações de tradução de proteínas trans-membrana e proteínas secretadas.

REAÇÕES BIDIRECIONAIS Algumas reações podem ser bidirecionais chegando a um equilíbrio dinâmico. Essas reações foram representadas por dois nós reação diferentes, um para cada sentido da reação. Muitas dessas reações são transporte trans-membrana, sendo a posição intra e extracelular representadas por dois nós molécula distintos.

REAÇÕES DE TRADUÇÃO A síntese de proteínas foi modelada em duas reações para cada proteína, sendo elas "*Translation Initiation Complex Formation*[3]" e "*Translation Reaction*[4]". A primeira consiste na formação do complexo pelo mRNA com a subunidade 30S do ribosomo. Foram ligados como modificadores à reação as moléculas RNA helicase, Fatores de iniciação de tradução 1, 2 e 3, fator de enlongamento (EF) P, tRNA com formil-metionina e GTPs como fontes de energia para o sistema. Como produto, um "*Translation Initiation Complex*[5]" (complexo IC) é criado, o qual é um reagente para o próximo passo. Na segunda reação agrupa todo o processo de polimerização e conformação tridimensional da proteína. Os reagentes são o com-

2 complexo de iniciação de replicação
3 Formação do Complexo de Iniciação de Tradução
4 Reação de Tradução
5 Complexo de Iniciamento de Tradução

plexo IC, grupos prostéticos e todos os tRNA-aminoácidos requeridos para a síntese proteica. Os modificadores dessa reação são as subunidades do ribossomo 30S e 50S, metionina deformylase e peptdase, EF-P, EF-4, fatores de liberação e chaperonas. Os produtos da reação de tradução são o monômero da proteína (ou monômeros se for o caso de um poly-mRNA) e os tRNA sem aminoácidos. Às reações de síntese de proteínas transmembrana e proteínas secretadas, também foram adicionados como modificadores proteínas de transporte, como a translocase, aminopeptdases extracelulares e partículas de reconhecimento de sinais (SRPs).

REAÇÕES DE INIBIÇÃO Como não existem links de inibição na rede proposta, as reações de inibição estão implicitamente representadas por reações as quais tem como produto um nó molécula diferente representando a molécula inibida. Por exemplo, a inativação de uma proteína por um ligante é dado pelo nó reação o qual liga a proteína e o ligante como reagentes e liga o nó da proteína inativada como um produto.

ANÁLISE TOPOLÓGICA

Diversas redes biológicas, como exemplificado na seção 2.3, compartilham uma mesma topologia chamada livre-de-escala. Estas redes representam subsistemas celulares os quais por si só apresentam as características de robustez e resiliência embasadas na topologia que as estrutura. Nesta linha de pensamento, levanta-se a questão: uma rede que represente uma célula completa também apresentaria esta topologia? Para responder esta questão, foi realizada a caracterização da topologia da rede gerada neste trabalho a qual modela todos os processos celulares conhecidos no organismo *Mycoplasma genitalium*.

Para analisar a topologia de uma rede bipartida, o método mais comum para extrair a distribuição de grau de um dos conjuntos separadamente é projetar a rede de um modo a partir da rede orginal. Contudo, gerar a projeção de um modo, utilizando o conjunto de nós molécula, nesta rede biológica implica em perder informação sobre as interações moleculares uma vez que a rede gerada se torna muito densa. Foi optado então por gerar a distribuição de grau da rede bipartida original, isolando somente os nós molécula para a análise topológica. Desta forma, o grau de um nó molécula relaciona o número de reações de que ele participa. Foi optado também por agrupar reações similares, como as de síntese protéica, para evitar nós com grau erroneamente elevado. Por exemplo, um ribossomo que está ligado à centenas de reações de síntese protéica, uma para cada RNA codificante do organismo, na prática, ele participa de somente uma reação, de síntese protéica, onde o substrato pode variar. de A Fig 16 apre-

senta a função de probabilidade acumulada de grau encontrada para links não direcionados, links de entrada e links de saída.

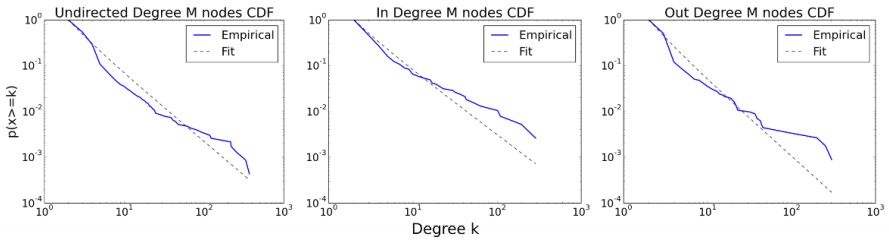


Figura 16: Função de probabilidade acumulada (CDF) dos nós molécula para links não direcionados, links de entrada e links de saída. As linhas pontilhadas indicam leis de potência que mais se aproximam das curvas empíricas com exponente α iguais à 2.47,2.39 e 2.65 respectivamente.

Pode-se observar que todas as curvas seguem um comportamento próximo a uma lei de potência com o expoente α entre 2 e 3 como observado na literatura. Para os links não direcionados, pode-se observar uma diferença da curva empírica para uma lei de potência ideal obtendo uma região de graus com probabilidade menor que o ideal. Firmando uma hipótese de que a curva empírica se aproximaria ao máximo de uma lei de potência, pode-se dizer que esta diferença obtida poderia ser corrigida aumentando o grau de nós com baixo grau, em outras palavras, nem todas as interações existentes estariam mapeadas na rede. Para verificar a completude da rede, podemos buscar por nós molécula que representem proteínas ou complexos protéicos e observar se todos tem funções mapeadas dentro da célula. Conforme o método utilizado para a modelagem da rede, todo nó molécula que representa uma proteína está conectada a pelo menos 3 reações, uma de síntese e duas de degradação, por proteases Lon e FtsH. Compelxos proteicos por sua vez tem seus nós molécula ligados a pelo menos 2 reações, uma de síntese e uma de degradação por protease Lon. Desta forma, buscando por nós molécula que representam proteínas e complexos proteicos com grau $k <= 3$ e $k <= 2$ respectivamente podemos encontrar proteínas e complexos protéicos sem função descrita na literatura, e portanto, não modelada na rede. Os nós encontrados nessa busca representam aproximadamente 20% das proteínas e complexos protéicos conhecidos neste organismo. Estas moléculas estão listadas na Tabela A.3 encontrada no Apêndice A.

MEDIDAS DA REDE

Para caracterizar a rede gerada neste trabalho, medidas de rede foram aplicadas para extrair informações sobre sua estrutura e organização.

A Tabela 1 descreve as medidas obtidas da *Whole-Cell Network* do organismo *Mycoplasma genitalium*.

Tabela 1: Métricas da *Whole-Cell Network* do organismo *Mycoplasma genitalium*

Medida	**Valor**
# de nós	6.630
# de links	85.354
# de componentes	1
Diâmetro	16
Média dos mínimos caminhos	5,13

PREDIÇÃO DE GENES ESSENCIAIS

Afim de testar as capacidades de modelagem da rede gerada neste trabalho, foi realizado um experimento para identificar genes essenciais para o organismo baseando-se somente na estrutura estática da rede. Para realizar este experimento, foi suposto que um nó molécula importante para a estrutura da rede idealmente representaria uma molécula importante para o organismo.

Pode-se dizer que uma molécula é importante para um organismo quando sem ela, o organismo não consegue realizar suas funções básicas de manutenção e replicação. Em outras palavras, metabólitos, proteínas, genes ou estruturas que quando sejam removidos do sistema, e assim deixando de existir as reações as quais eles participam, levem o organismo à morte ou à não reprodução, são considerados importantes. O organismo estudado nesse trabalho, *Mycoplasma genitalium*, é alvo de diversas pesquisas sobre componentes moleculares mínimos para que se haja vida, sendo ele o protagonista por ser detentor do menor genoma conhecido até então. Sendo assim, encontrar o conjunto mínimo de genes essenciais para a manutenção e reprodução da célula deste organismo é o alvo deste experimento.

Dado que para uma reção bioquímica acontecer, todos seus reagentes precisam estar presentes no mesmo meio. A remoção de uma molécula de um sistema implica que imediatamente, todas as reações as quais essa molécula participava como reagente deixam de existir. Por indução, os produtos desta reação deixam de ser produzidos por este meio, sendo assim, caso este produto não seja também produzido por uma reação distinta, ele também deixa de existir no sistema. Consequentemente, a remoção de uma molécula em um determinado sistema pode acarretar uma cascata de remoção de outras moléculas por indução. Isto ocorre naturalmente em células quando a inibição do funcionamento de uma proteína acarreta em uma sequência de outras inibições, acontecendo frequentemente em regulagem

de sistemas fisiológicos. Também ocorre de forma artificial quando um antibiótico inibe, dentro de uma bactéria, o funcionamento de uma molécula, acarretando em uma cascata de falhas as quais levam o organismo à morte.

De forma análoga, a remoção de um nó molécula da rede gerada implica na remoção dos nós reação os quais o nó molécula participa como reagente. Dentro desse conjunto de nós reação removidos do sistema, se algum nó mólecula produto dessas reações passar a ter um grau de entrada igual a zero, ele também é removido do sistema iniciando o processo novamente. O algorítmo 1 descreve o processo de remoção em cascata de nós dado a remoção de um nó v.

Algorithm 1 Remoção de nós em cascata

```
1:  procedure CASCADENODEREMOVAL(N)   ▷ N is the molecule node to remove
2:      R ← list of actions where N is reactant
3:      M ← empty list of molecules
4:      C ← 0                                   ▷ cascade steps counter
5:      RNC ← Number of molecule nodes in the Network   ▷ Remaining Nodes' Count
6:      remove(N)
7:      RNC ← RNC − 1
8:      while Length(R) > 0 do
9:          for all r in R do
10:             remove(r)
11:             for all reactant of r do
12:                 if reactant.indegree = 0 then
13:                     append reactant to M
14:         R ← empty list of reactions
15:         for all m in M do
16:             remove(m)
17:             RNC ← RNC − 1
18:             append reactions where m is reactant to R
19:             M ← empty list of molecules
20:         C ← C + 1
        return RNC, C
```

Uma vez que a rede gerada neste trabalho visa representar o sistema completo de um organismo, foi feita a seguinte analogia: um nó molécula importante para a estrutura da rede deve representar uma molécula importante para o organismo. Desta forma, classificamos um nó molécula importante para a rede como um nó cuja sua remoção e a remoção dos nós reação subsequentes os quais o nó molécula participa como reagente não impacte a quantidade total de nós da rede em mais que uma porcentagem definida empiricamente.

Como primeiro passo do experimento, cada nó molécula foi removido individualmente do sistema utilizando o Algoritmo 1 e anali-

zado seu impacto na quantidade de nós molécula total remanascentes na rede após a cascata de remoção. A Fig 17a mostra na linha azul a quantidade de nós remanescentes na rede (RNC - *Remaining Nodes Count*) após a remoção individual dos nós molécula. Cada ponto no eixo horizontal representa um nó molécula dos 2740 nós ordenados de forma decrescente pelo seu RNC. A linha vermelha mostra a quantidade de passos da cascata de remoção.

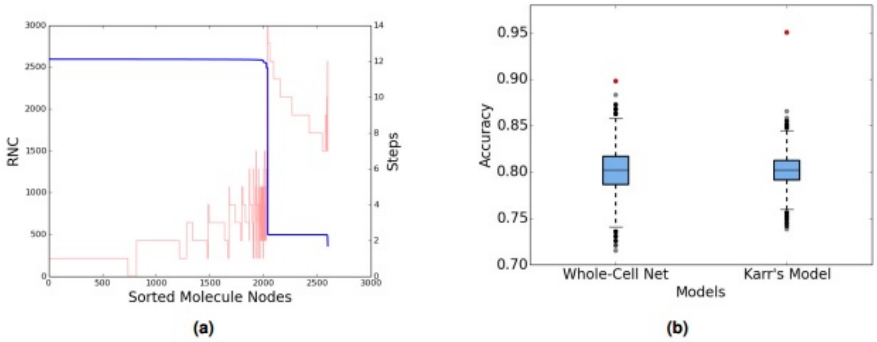


Figura 17: **a)** Em azul, os RNCs ordenados de forma decrescente de cada nó molécula da rede. Em vermelho, o número de passos da cascata de remoção de nós. **b)** Boxplot da simulação estatística com 10 mil amostras aleatórias para o modelo *Whole-Cell Network* e modelo de Karr.

Pode-se observar também uma queda brusca no valor de RNC ao longo da curva. Um baixo RNC indica um alto impacto na rede pela remoção do nó correspondente e subsequente cascata de remoção. No limiar da queda brusca da curva, podemos observar dois grupos distintos de moléculas. As moléculas com alto RNC indicam moléculas de baixo impacto na rede formando um grupo e outro grupo com RNC $\leqslant$ 498 os quais demonstram um elevado impacto na rede. Este segundo grupo de moléculas de alto impacto na estrutura da rede foram consideradas como cruciais à estrutura da rede e, por consequência, cruciais para o funcionamento e sobrevivência do organismo. Dentre estas moléculas, foram identificados 140 genes cruciais à célula.

Em um segudo momento, foram removidos nós da rede sequencialmente ordenados pelo seu RNC de forma decrescente, realizando para cada um a cascata de remoção descrita anteriormente. A remoção e cascata de remoção somente se mantinha caso a quantidade total de nós da rede, após a remoção em cascata, fosse reduzido em no máximo 5%. Caso contrário, o nó removido e todos os nós removidos na cascata eram restabelecidos na rede novamente. O algorítmo 2 descreve o processo de remoção sequencial dos nós da rede.

Ao final deste processo, os 755 nós molécula interligados por 609 nós reação remanescentes na rede foram considerados como os que representam moléculas essenciais à vida do organismo. Dentre os nós molécula remanescentes, foram selecionados os que representam ge-

Algorithm 2 Remoção em cascata sequencial

```
1: procedure SEQUENTIALCASCADEREMOVAL(Network)
2:     for all molecule node N_i in Network do
3:         RNC_i, C_i ← cascadeNodeRemoval(N_i)
4:     sort N by RNC and secondarily by C
5:     reverse order of N
6:     for all molecule node V_i in Network do
7:         if N_i is in Network then
8:             g ← Number of molecule nodes in Network
9:             new_g, c ← cascadeNodeRemoval(N_i)
10:            if new_g < 0.95 * g then
11:                undo cascadeNodeRemoval(N_i)
       return Network
```

nes e esse conjunto de 197 genes, o qual consideramos essencial para o organismo, foi comparado com cinco conjuntos de genes essenciais para o mesmo organismo obtidos na literatura. Dentre estes conjuntos se encontram o conjunto gerado por simulação pelo modelo computacional de Karr para simulação de células completas[18], o conjunto gerado experimentalmente por Glass utilizando transferência aleatória de transposons[43], o conjunto gerado pelo modelo teórico por ortólogos de B. subtilis por Kobayashi[44], o conjunto gerado pelo modelo teórico de Mushegian & Koonin[45] e o conjunto teórico de Gil[**?**]. A Tabela 2 mostra a correspondência entre o conjunto gerado por este trabalho e os conjuntos encontrados na literatura. A acurácia do conjunto de genes essenciais encontrados neste trabalho foi medida pela razão entre número de genes considerados essenciais que concordam com o modelo comparado e o total de genes considerados essenciais.

Tabela 2: Comparação com outros modelos de genes essenciais preditos.

	Karr's Model	**Glass Set**	***B. subtilis* orthologs**	**Mushegian & Koonin Set**	**Gil Core Set**
Correspondências	379	261	389	300	375
Falsos Positivos	29	20	85	41	71
Falsos Negativos	113	240	51	184	79
Acurácia	85.28%	89.85%	56.85%	79.19%	63.96%

Para confirmar a relevância estatística deste resultado, foi utilizado como referência o conjunto de genes essenciais determinados experimentalmente por Glass e comparado com o modelo computacional de Karr. Foram gerados 10.000 conjuntos aleatórios com o mesmo número de genes do conjunto de Glass dentre o total de 525 genes do

organismo. Com cada um dos conjuntos aleatórios foi determinada a porcentagem de correspondências entre os genes essenciais do conjunto gerado por esse trabalho e o conjunto gerado pelo modelo computacional de Karr. A Figura 17b demonstra no boxplot o espectro de correspondência com os conjutos aleatórios e no ponto vermelho a correspondência com o conjunto real de genes essenciais.

4

CONCLUSÕES

Neste estudo foi proposto um novo modelo de rede chamado *Whole-Cell Network*, que pode representar de uma maneira totalmente integrada processos celulares distintos, os quais eram previamente modelados utilizando diferentes modelos de rede. Foi desenvolvido um método consistente para a representação de organismos inteiros, com o objetivo de contribuir para avanços recentes em modelos de simulação de células completas. A *Whole-Cell Network* é constituída por nós molécula e nós reação. Nós molécula podem representar quaisquer moléculas ou estruturas na célula enquanto nós reação representam interações entre nós molécula. Os links que ligam os nós molécula aos nós reação são ponderados pela estequiometria da reação. Catalisadores, tais como enzimas, estão ligados a nós reação como modificadores. Portanto, todas as interações moleculares que ocorrem no interior de uma célula podem ser representadas por uma única rede homogênea. Além disso, a *Whole-Cell Network* pode fornecer dados estruturados e ferramentas para melhor compreender o papel de moléculas particulares num contexto sistêmico.

Foi modelado como um estudo de caso a primeira *Whole-Cell Network* de um organismo. O organismo escolhido foi o *Mycoplasma genitalium*, bactéria patogênica, a qual já foi objeto de diversos estudos de modelagem por causa da sua relativa simplicidade. Utilizando dados de larga escala deste organismo, foi gerado um modelo que pôde integrar vários processos celulares, incluindo: replicação de DNA, danos no DNA, reparo de DNA, citocinese, expressão de gênica, degradação de RNA, síntese de proteínas, metabolismo, interações proteína-proteína, formação de complexos protéicos, degradação de proteínas e transporte. A rede resultante se mostrou composta por uma única componente gigante, isto é, todas as moléculas e interações representados no modelo puderam ser ligados uns aos outros utilizando um número finito de conexões. Era esperado essa característica surgir naturalmente a partir destes dados de larga escala uma vez que fossem suficientes para representar todos os processos celulares. Caso contrário, se não fosse alcançado uma única componente, poderíamos inferir insuficiência de dados relativos aos processos do organismo.

A análise topológica da *Whole-Cell Network* do *M. genitalium* indicou que esta rede tem uma distribuição de grau que é compartilhada por muitas redes biológicas. Esta topologia, chamada livre-de-escala, está subjacente a vários sistemas naturais e pensa-se ser responsável por muitas das características de células, incluindo robustez e resiliência. O elevado número de nós com poucas conexões indica a baixa pro-

babilidade de uma falha aleatória nos nós com um elevado número de ligações (conhecido como hubs). Esta característica reduz a probabilidade de falha em moléculas essenciais para a célula, tais como a mutação genética em enzimas importantes ou ausência de nutrientes essenciais. Por outro lado, esta é uma ferramenta poderosa para identificar os nós essenciais como pontos fracos de uma célula, os quais podem ser alvos de drogas em organismos patogênicos.

A análise métrica da rede gerada indica características de redes de mundo pequeno, onde mesmo com um alto número de nós, em média pode-se chegar de um nó a outro em poucos passos. Considerando que esta é uma rede bipartida, na prática o caminho de interações físicas entre duas moléculas se dá pela metade do caminho medido na rede, onde reações também são consideradas como nós. O valor do mínimo caminho médio cai para aproximadamente 2,5, valor o qual mostra a proximidade de relações físicas entre as moléculas de uma célula.

Foram realizadas duas experiências para testar a hipótese de que moléculas importantes para a sobrevivência da célula são representadas por nós importantes para a estrutura da rede. Ao remover individualmente cada nó, medimos impacto da sua remoção sobre a estrutura de rede. Como esperado, moléculas tais como ATP, GTP, aminoácidos e proteínas tais como ribossomos e polimerases, obtiveram um impacto significativo sobre a rede. Outras moléculas, cuja essencialidade não era tão óbvia, incluindo a co-chaperona GrpE, também puderam ser observadas com um impacto elevado. No entanto, esta experiência tem algumas limitações. Em primeiro lugar, a *Whole-Cell Network* é uma representação estática da célula e representa todas as reações possíveis que podem ocorrer em qualquer momento ou cenário. Portanto, as moléculas importantes cuja ausência indiretamente levaria à morte celular não puderam ser identificadas. Em segundo lugar, a essencialidade dos genes do metabolismo pode variar a depender da disponibilidade de nutrientes. No segundo experimento, nós podamos a rede removendo nós sequencialmente para alcançar uma rede mínima essencial onde os nós molécula restantes foram classificados como essenciais para a célula. A comparação entre nós moléculas que representam genes pela nossa classificação e genes essenciais obtidos a partir de dados experimentais indicou uma alta precisão de classificação. No que diz respeito aos falsos negativos, podem ser associados as mesmas limitações da primeira experiência, pois é um modelo estático e não pode capturar aspectos dinâmicos da célula. Apesar de sua natureza estática, esta representação classificou corretamente 16 genes como essenciais os quais não foram corretamente classificados pelo modelo computacional de Karr o qual é dinâmico.

Embora os dados biológicos de grande escala estão amplamente disponíveis atualmente, os métodos para integrar estes dados de forma

consistente são limitados. Representações disjuntas de rede ainda não conseguem capturar as inter-relações entre processos celulares. Por esta razão, acreditamos que este estudo pode contribuir para uma representação consistente de organismos inteiros e fornecer dados úteis para futuros estudos envolvidos em simulações de células completas.

A

APÊNDICE

FORMATO DE ARQUIVOS

O *M. Genitalium Whole-Cell Network* foi gerado e armazenado em dois formatos, SBML e GML.

MG_NETWORK.SBML Os nós molécula são armazenados como objetos Species, nós reação como objetos Reaction. A representação comum da reação de modelos de SBML é usada mas sem leis cinéticas associadas. Moléculas catalíticas como enzimas são conectadas a reações como modificadores. O tipo de moléculas e reações são armazenadas como anotações e processos celulares de que as reações são parte de são armazenados como notas.

MG_NETWORK.GML O formato de grafo comum que armazena nós e links com atributos associados. As moléculas e reações são representadas por nós com valor diferente no atributo 'type' (0 para nós molécula e 1 para nós reação. Todos os atributos de nós e links são descritos abaixo.

Atributos de nós:

- name: string ID único para cada nó. O campo 'wid' no WholeCellKB database;
- bdName: Nome de nó estendido com função de molécula para nós molécula;
- type: 0 para nós molécula e 1 para nós reação;
- cellLocation: um único caractere para identificar localização do nó molécula. 'c': cytosol; 'm': membrane; 'e': extracellular;
- annotation: tipo da molécula ou reação;
- process: processo de qual faz parte;
- Degree: número de conexões do nó.

Atributos dos links:

- type: tipo de conecção entre nós molécula e nós reação (reactant, product, modifier);
- weight: estoquiometria.

REPLICABILIDADE DO DATASET

Para tornar este conjunto de dados reproduzível, disponibilizamos todos os scripts desenvolvidos para construir a rede caso requisitados. Existem dois scripts Python 2.7, um para ler o banco de dados e criar o modelo em formato SBML e outro para converter o arquivo SBML em um arquivo GML. O uso de cada um é descrito abaixo.

BIONETWORK_MYSQL_TO_SBML.PY Este script requer a instalação prévia do compilador Python 2.7, biblioteca SBML API e MySQLdb lib para Python e EMBOSS transeq executando localmente. Deve estar funcionando também um banco de dados MySQL com o *Knowledge Database* do *Mycoplasma genitalium*. Este banco de dados pode ser obtido em `https://simtk.org/frs/download.php?file_id=3426` com o arquivo `data.sql` dentro do arquivo zip. Para executar o script, use o comando "`python bionetwork_mysql_to_sbml.py -h`" e as instruções para executar serão exibidas.

CONVERTER_SBML_TO_GML.PY Este script requer a instalação prévia do compilador Python 2.7, SBML API lib e igraph lib para Python. Para executar o script, use o comando "`python converter_SBML_to_GML.py -h`" e as instruções para executar serão exibidas.

O primeiro script só pode ser aplicado diretamente a esse banco de dados específico por causa de alguns nomes de molécula inseridos manualmente no código. No entanto, o código pode ser adaptado para banco de dados de outro organismo caso disponível. O segundo script pode ser aplicado a qualquer arquivo SBML gerado pelo primeiro.

PROTEÍNAS COM FUNÇÃO DESCONHECIDA

A *Whole-Cell Network* do *Mycoplasma genitalium* foi modelada de tal forma que cada nó molécula que representa um monômero de proteína tem grau mínimo de 3 (uma reação de síntese e duas reações de degradação pelas proteases FtsH e Lon) e nós molécula que representam complexos protéicos têm grau de no mínimo 2 (reação de síntese e reação de degradação pela protease Lon). Deste modo, monômeros de proteínas e complexos de proteínas com os respectivos graus podem ser classificados como moléculas sem função conhecida por não participarem em outras reações. A Tabela A.3 lista os monómeros ou complexos de proteínas que se encaixam no grau de nó descrito.

Tabela 3: Proteínas com função desconhecida

Nó	Tipo	Nome
MG_002_MONOMER	Monomer	DnaJ domain protein
MG_010_MONOMER	Monomer	DNA primase-related protein
MG_011_MONOMER	Monomer	
MG_018_MONOMER	Monomer	SNF2 family helicase putative
MG_024_MONOMER	Monomer	GTP-binding protein YchF
MG_025_MONOMER	Monomer	Beta-glycosyl transferase
MG_028_MONOMER	Monomer	
MG_032_MONOMER	Monomer	
MG_474_MONOMER	Monomer	
MG_056_MONOMER	Monomer	tetrapyrrole (corrin/porphyrin) methylase protein
MG_057_MONOMER	Monomer	small primase-like protein
MG_060_MONOMER	Monomer	glycosyl transferase
MG_067_MONOMER	Monomer	lipoprotein, putative
MG_068_MONOMER	Monomer	lipoprotein, putative
MG_074_MONOMER	Monomer	
MG_075_MONOMER	Monomer	116 kDa surface antigen
MG_076_MONOMER	Monomer	
MG_095_MONOMER	Monomer	lipoprotein, putative
MG_096_MONOMER	Monomer	
MG_103_MONOMER	Monomer	
MG_108_MONOMER	Monomer	Ser/Thr protein phosphatase 2C, putative
MG_115_MONOMER	Monomer	competence/damage-inducible protein CinA domain protein
MG_116_MONOMER	Monomer	
MG_117_MONOMER	Monomer	
MG_125_MONOMER	Monomer	Cof-like hydrolase, putative
MG_129_MONOMER	Monomer	Putative phosphotransferase enzyme glucose-specific IIB component
MG_131_MONOMER	Monomer	hypothetical protein
MG_133_MONOMER	Monomer	membrane protein, putative

Continua na próxima página

Tabela 3 – *Continuação da página anterior*

Nó	Tipo	Nome
MG_134_MONOMER	Monomer	
MG_135_MONOMER	Monomer	membrane protein, putative
MG_140_MONOMER	Monomer	putative ATP-dependent helicase
MG_477_MONOMER	Monomer	
MG_144_MONOMER	Monomer	
MG_146_MONOMER	Monomer	UPF0053 protein
MG_147_MONOMER	Monomer	membrane protein, putative
MG_148_MONOMER	Monomer	
MG_149_MONOMER	Monomer	lipoprotein, putative
MG_478_MONOMER	Monomer	
MG_185_MONOMER	Monomer	lipoprotein, putative
MG_199_MONOMER	Monomer	ribonuclease HIII, putative, frameshift
MG_200_MONOMER	Monomer	DnaJ domain protein
MG_202_MONOMER	Monomer	
MG_207_MONOMER	Monomer	Ser/Thr protein phosphatase 2A
MG_480_MONOMER	Monomer	
MG_211_MONOMER	Monomer	
MG_491_MONOMER	Monomer	
MG_219_MONOMER	Monomer	
MG_220_MONOMER	Monomer	
MG_222_MONOMER	Monomer	S-adenosyl-methyltransferase MraW
MG_223_MONOMER	Monomer	
MG_233_MONOMER	Monomer	
MG_237_MONOMER	Monomer	
MG_241_MONOMER	Monomer	
MG_242_MONOMER	Monomer	
MG_243_MONOMER	Monomer	conserved hypothetical protein, authentic frameshift
MG_246_MONOMER	Monomer	Ser/Thr protein phosphatase 2A
MG_247_MONOMER	Monomer	membrane protein, putative
MG_248_MONOMER	Monomer	
MG_255_MONOMER	Monomer	

Continua na próxima página

Tabela 3 – *Continuação da página anterior*

Nó	**Tipo**	**Nome**
MG_494_MONOMER	Monomer	
MG_256_MONOMER	Monomer	
MG_260_MONOMER	Monomer	lipoprotein, putative
MG_263_MONOMER	Monomer	Cof-like hydrolase
MG_267_MONOMER	Monomer	
MG_268_MONOMER	Monomer	
MG_269_MONOMER	Monomer	
MG_279_MONOMER	Monomer	
MG_280_MONOMER	Monomer	sensory rhodopsin II transducer motif
MG_281_MONOMER	Monomer	
MG_284_MONOMER	Monomer	
MG_285_MONOMER	Monomer	
MG_286_MONOMER	Monomer	
MG_288_MONOMER	Monomer	protein L
MG_505_MONOMER	Monomer	putative holliday junction resolvase
MG_294_MONOMER	Monomer	major facilitator superfamily (MFS) protein
MG_296_MONOMER	Monomer	
MG_306_MONOMER	Monomer	membrane protein, putative
MG_307_MONOMER	Monomer	lipoprotein, putative
MG_309_MONOMER	Monomer	lipoprotein, putative
MG_313_MONOMER	Monomer	membrane protein, putative
MG_314_MONOMER	Monomer	
MG_319_MONOMER	Monomer	
MG_320_MONOMER	Monomer	membrane protein, putative
MG_515_MONOMER	Monomer	
MG_326_MONOMER	Monomer	degV family protein
MG_328_MONOMER	Monomer	coiled coil putative structural protein involved in cytoskeleton
MG_331_MONOMER	Monomer	
MG_332_MONOMER	Monomer	
MG_516_MONOMER	Monomer	
MG_337_MONOMER	Monomer	

Continua na próxima página

Tabela 3 – *Continuação da página anterior*

Nó	Tipo	Nome
MG_338_MONOMER	Monomer	lipoprotein, putative
MG_343_MONOMER	Monomer	
MG_348_MONOMER	Monomer	lipoprotein, putative
MG_350_MONOMER	Monomer	
MG_521_MONOMER	Monomer	membrane protein, putative
MG_354_MONOMER	Monomer	
MG_360_MONOMER	Monomer	ImpB/MucB/SamB family protein
MG_366_MONOMER	Monomer	
MG_371_MONOMER	Monomer	DHH family protein
MG_373_MONOMER	Monomer	
MG_374_MONOMER	Monomer	
MG_377_MONOMER	Monomer	
MG_381_MONOMER	Monomer	
MG_524_MONOMER	Monomer	
MG_388_MONOMER	Monomer	Putative metalloprotease
MG_389_MONOMER	Monomer	
MG_395_MONOMER	Monomer	lipoprotein, putative
MG_397_MONOMER	Monomer	
MG_406_MONOMER	Monomer	
MG_414_MONOMER	Monomer	
MG_525_MONOMER	Monomer	
MG_422_MONOMER	Monomer	
MG_423_MONOMER	Monomer	metallo-beta-lactamase superfamily protein
MG_432_MONOMER	Monomer	membrane protein, putative
MG_439_MONOMER	Monomer	lipoprotein, putative
MG_440_MONOMER	Monomer	lipoprotein, putative
MG_441_MONOMER	Monomer	
MG_443_MONOMER	Monomer	membrane protein, putative
MG_449_MONOMER	Monomer	conserved hypothetical protein, authentic frameshift
MG_450_MONOMER	Monomer	degV family protein
MG_452_MONOMER	Monomer	membrane protein, putative
MG_456_MONOMER	Monomer	

Continua na próxima página

Tabela 3 – *Continuação da página anterior*

Nó	**Tipo**	**Nome**
MG_459_MONOMER	Monomer	putative 2-C-methyl-D-erythritol-2,4-cyclodiphosphate synthase
MG_461_MONOMER	Monomer	HD domain protein
MG_029_DIMER	Complex	DJ-1/PfpI family protein
MG_064_065_TETRAMER	Complex	ABC transporter
MG_085_HEXAMER	Complex	HPr(Ser) kinase/phosphatase
MG_102_DIMER_ox	Complex	oxidized thioredoxin-disulfide reductase
MG_127_MONOMER_ox	Complex	oxidized Spx subfamily protein
MG_132_DIMER	Complex	purine nucleoside phosphoramidase
MG_208_DIMER	Complex	glycoprotease family protein
MG_271_272_273_274_192MER_ox	Complex	oxidized dihydrolipoamide dehydrogenase
MG_295_MONOMER_ox	Complex	oxidized tRNA U34 sulfurtransferase
MG_353_DIMER	Complex	DNA-binding protein HU, putative
MG_376_HEXAMER	Complex	
MG_409_DIMER	Complex	phosphate transport system regulatory protein PhoU, putative

REFERÊNCIAS BIBLIOGRÁFICAS

[1] G. B., Rene Descartes, and Laurence J. Lafleur. Discourse on Method. *The Journal of Philosophy*, 47(16): 473, 1950. ISSN 0022362X. doi: 10.2307/2021377. URL http://www.pdcnet.org/oom/service?url{_}ver=Z39.88-2004{&}rft{_}val{_}fmt={&}rft.imuse{_}id=jphil{_}1950{_}0047{_}0016{_}0473{_}0474{&}svc{_}id=info:www.pdcnet.org/collection.

[2] F Capra. Ponto de Mutação, O. *SP: Cultrix*, pages 284–337, 1982.

[3] P a Weiss. *Within the Gates of Science and Beyond: Science in ist Cultural Commitments*. New York, Hafner Pub. Co., 1971.

[4] Colin Macilwain. Systems biology: Evolving into the mainstream. *Cell*, 144(6):839–841, mar 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.02.044. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867411002340.

[5] H. Kitano. Systems Biology: A Brief Overview. *Science*, 295 (5560):1662–1664, mar 2002. ISSN 00368075. doi: 10.1126/science.1069492. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1069492.

[6] Béla Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*, 286(1):257–257, 1984. ISSN 0002-9947. doi: 10.1090/S0002-9947-1984-0756039-5.

[7] Frederic Amblard. *Small Worlds: The Dynamics of Networks between Order and Randomness*, volume 6. 2003. ISBN 0691005419. doi: 10.1146/annurev.soc.30.020404.104342. URL http://books.google.com/books?id=rydKGwfs3UAC.

[8] M Covert. Regulation of Gene Expression in Flux Balance Models of Metabolism. *Journal of Theoretical Biology*, 213 (1):73–88, 2001. ISSN 00225193. doi: 10.1006/jtbi.2001.2405. URL citeulike-article-id$\delimiter"026E30F$n500934$\delimiter"026E30F$nhttp$\delimiter"026E30F$n//dx.doi.org/10.1006/jtbi.2001.2405.

[9] Katsuyuki Yugi, Hiroyuki Kubota, Yu Toyoshima, Rei Noguchi, Kentaro Kawata, Yasunori Komori, Shinsuke Uda, Katsuyuki Kunida, Yoko Tomizawa, Yosuke Funato, Hiroaki Miki, Masaki Matsumoto, Keiichi I. Nakayama, Kasumi Kashikura, Keiko Endo,

Kazutaka Ikeda, Tomoyoshi Soga, and Shinya Kuroda. Reconstruction of insulin signal flow from phosphoproteome and metabolome data. *Cell Reports*, 8(4):1171–1183, 2014. ISSN 22111247. doi: 10.1016/j.celrep.2014.07.021. URL http://dx.doi.org/10.1016/j.celrep.2014.07.021.

[10] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. Getting connected: Analysis and principles of biological networks. *Genes and Development*, 21(9):1010–1024, 2007. ISSN 08909369. doi: 10.1101/gad.1528707. URL http://www.genesdev.org/cgi/doi/10.1101/gad.1528707.

[11] Albert-Laszlo Barabasi, Zoltan N. Zoltán N Oltvai, and Albert-László Barabási. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. ISSN 1471-0056. doi: 10.1038/nrg1272. URL http://www.ncbi.nlm.nih.gov/pubmed/14735121.

[12] Hawoong Jeong, B Tombor, Réka Albert, Zoltán N Oltvai, and Albert-László Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, oct 2000. ISSN 0028-0836. doi: 10.1038/35036627. URL http://www.hubmed.org/display.cgi?uids=11034217.

[13] Réka Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947–4957, 2005. ISSN 0021-9533. doi: 10.1242/jcs.02714. URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.02714.

[14] Raya Khanin and Ernst Wit. How Scale-Free Are Biological Networks. *Journal of Computational Biology*, 13(3):810–818, 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.810. URL http://www.liebertonline.com/doi/abs/10.1089/cmb.2006.13.810.

[15] Timothy G Buchman. The community of the self. *Nature*, 420 (6912):246–251, 2002. ISSN 0028-0836. doi: 10.1038/nature01260.

[16] M Vidal, M E Cusick, and A L Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, mar 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.02.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867411001309.

[17] D K Arrell and a Terzic. Network systems biology for drug discovery. *Clinical pharmacology and therapeutics*, 88(1):120–125, 2010. ISSN 1532-6535. doi: 10.1038/clpt.2010.91.

[18] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. MacKlin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, jul 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.

05.044. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412007763.

[19] Mark Newman. *Networks: An Introduction*. 2010. ISBN 9780191594175. doi: 10.1093/acprof:oso/9780199206650.001.0001. URL http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199206650.001.0001/acprof-9780199206650.

[20] B.˜Bollobás. Random Graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1985. ISSN 0003-4851. doi: 10.1214/aoms/1177706098. URL http://books.google.com/books?hl=en{&}lr={&}id=o9WecWgilzYC{&}oi=fnd{&}pg=PR10{&}dq=Random+Graphs{&}ots=YzLPmVUrQn{&}sig=ZKaH65n3j1jNXpM2b4Aqug6f-a8$\delimiter"026E30F$nhttp://books.google.com/books?hl=en{&}lr={&}id=o9WecWgilzYC{&}oi=fnd{&}pg=PR10{&}dq=Random+graphs{&}ots=YzLPmVUrRk{&}sig=zZck0IzC1Tex.

[21] Jeffrey Travers and Stanley Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425–443, 1969. ISSN 0038-0431. doi: 10.2307/2786545.

[22] Ronald S. Burt. *Structural holes: The social structure of competition*. 2009. ISBN 9780674843714. URL http://books.google.com/books?hl=en{&}lr={&}id=FAhiz9FWDzMC{&}oi=fnd{&}pg=PR5{&}dq={%}22These+three+kinds+of+data+are+transformed+in+various+ways{%}22+{%}22the+lack+of+a+structural+hole,+and+cohesion+is+the+more{%}22+{%}22redundant+to+the+extent+that:+(a)+you+have+a+substant.

[23] Stephen P. Borgatti. Structural holes: Unpacking Burt's redundancy measures. *Connections*, 20:35–38, 1997.

[24] Santo Fortunato. Community detection in graphs, 2010. ISSN 03701573.

[25] Matthew B Dobbs. Genetics in orthopaedics., 2007. ISSN 0009-921X. URL http://www.nature.com/physics/looking-back/crick/$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/13054692$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/17804964.

[26] Ph.D. Leslie A. Pray. Discovery of DNA Double Helix: Watson and Crick, 2008. ISSN 1570-0232. URL http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397.

[27] Frances Crick. Central Dogma of Molecular Biology, 1970. ISSN 0028-0836. URL http://dx.doi.org/10.1038/227561a0.

[28] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1):D459–D471, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1103. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1103.

[29] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, 1999. ISSN 03051048. URL http://nar.oxfordjournals.org/cgi/content/long/27/1/29.

[30] Hyun Uk Kim and Sang Yup Lee. Applications of genome-scale metabolic network models in the biopharmaceutical industry. *Pharmaceutical Bioprocessing*, 1(4):337–339, 2013. ISSN 2048-9145. doi: 10.4155/pbp.13.37. URL http://dx.doi.org/10.4155/pbp.13.37.

[31] Jeffrey D Orth, Ines Thiele, and B O Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010. ISSN 15461696. doi: 10.1038/nbt.1614. URL http://www.nature.com/doifinder/10.1038/nbt.1614.

[32] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261, 2003. ISSN 03051048. doi: 10.1093/nar/gkg034.

[33] Mike Tyers, Ashton Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, and Bobby-Joe Breitkreutz. BioGRID: a general repository for interaction datasets. *Nucl. Acids Res.*, 34(suppl_1):D535–539, 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj109. URL http://nar.oxfordjournals.org/cgi/content/abstract/34/suppl{_}1/D535$\delimiter"026E30F$nhttp://nar.oxfordjournals.org/cgi/reprint/34/suppl{_}1/D535.pdf.

[34] Amanda L Garner and Kim D Janda. Protein-protein interactions and cancer: targeting the central dogma. *Current topics in medicinal chemistry*, 11(3):258–280, 2011. ISSN 15680266. doi: 10.2174/156802611794072614.

[35] Patricia A J Muller and Karen H. Vousden. Mutant p53 in cancer: New functions and therapeutic opportunities, 2014. ISSN 18783686.

[36] David-Emlyn Parfitt and Michael M Shen. From blastocyst to gastrula: gene regulatory networks of embryonic stem cells and early mouse embryogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369 (1657), 2014. ISSN 1471-2970. doi: 10.1098/rstb.2013.0542. URL http://www.ncbi.nlm.nih.gov/pubmed/25349451$\delimiter"026E30F$nhttp://rstb.royalsocietypublishing.org/content/369/1657/20130542.abstract.

[37] S C Materna and P Oliveri. A protocol for unraveling gene regulatory networks. *Nat. Protoc.*, 3(12):1876–1887, 2008. ISSN 1754-2189. doi: nprot.2008.187[pii]$\backslash$r10.1038/nprot.2008.187. URL http://www.ncbi.nlm.nih.gov/pubmed/19008874$\delimiter"026E30F$nhttp://www.nature.com/nprot/journal/v3/n12/pdf/nprot.2008.187.pdf.

[38] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology*, 9(10): 770–780, 2008. ISSN 1471-0072. doi: 10.1038/nrm2503.

[39] Sunil Sethi, Gagandeep Singh, Palash Samanta, and Meera Sharma. Mycoplasma genitalium: An emerging sexually transmitted pathogen. *Indian Journal of Medical Research*, 136(6):942–955, 2012. ISSN 09715916. doi: 10.1016/j.medmal.2012.05.006. URL http://www.ncbi.nlm.nih.gov/pubmed/22975074.

[40] Clyde A Hutchison, Hamilton O Smith, and J Craig Venter. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science (New York, N.Y.)*, 2165(1999):2165–2170, 2012. ISSN 00368075. doi: 10.1126/science.286.5447.2165.

[41] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Abhishek Arora, and Markus W. Covert. WholeCellKB: Model organism databases for comprehensive whole-cell models. *Nucleic Acids Research*, 41(D1):D787–92, jan 2013. ISSN 03051048. doi: 10.1093/nar/gks1108. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531061{&}tool=pmcentrez{&}rendertype=abstract.

[42] Mickael Goujon, Hamish McWilliam, Weizhong Li, Franck Valentin, Silvano Squizzato, Juri Paern, and Rodrigo Lopez. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(SUPPL. 2):W695–W699, 2010. ISSN 03051048. doi: 10.1093/nar/gkq313. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq313.

[43] John I Glass, Nacyra Assad-Garcia, Nina Alperovich, Shibu Yooseph, Matthew R Lewis, Mahir Maruf, Clyde a Hutchison, Hamilton O Smith, and J Craig Venter. Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of*

the United States of America, 103(2):425–430, January 2006. ISSN 0027-8424. doi: 10.1073/pnas.0510013103.

[44] K Kobayashi, S D Ehrlich, a Albertini, G Amati, K K Andersen, M Arnaud, K Asai, S Ashikaga, S Aymerich, P Bessieres, F Boland, S C Brignell, S Bron, K Bunai, J Chapuis, L C Christiansen, a Danchin, M Débarbouille, E Dervyn, E Deuerling, K Devine, S K Devine, O Dreesen, J Errington, S Fillinger, S J Foster, Y Fujita, a Galizzi, R Gardan, C Eschevins, T Fukushima, K Haga, C R Harwood, M Hecker, D Hosoya, M F Hullo, H Kakeshita, D Karamata, Y Kasahara, F Kawamura, K Koga, P Koski, R Kuwana, D Imamura, M Ishimaru, S Ishikawa, I Ishio, D Le Coq, a Masson, C Mauël, R Meima, R P Mellado, a Moir, S Moriya, E Nagakawa, H Nanamiya, S Nakai, P Nygaard, M Ogura, T Ohanan, M O'Reilly, M O'Rourke, Z Pragai, H M Pooley, G Rapoport, J P Rawlins, L a Rivas, C Rivolta, a Sadaie, Y Sadaie, M Sarvas, T Sato, H H Saxild, E Scanlan, W Schumann, J F M L Seegers, J Sekiguchi, a Sekowska, S J Séror, M Simon, P Stragier, R Studer, H Takamatsu, T Tanaka, M Takeuchi, H B Thomaides, V Vagner, J M van Dijl, K Watabe, a Wipat, H Yamamoto, M Yamamoto, Y Yamamoto, K Yamane, K Yata, K Yoshida, H Yoshikawa, U Zuber, and N Ogasawara. Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8):4678–4683, 2003. ISSN 0027-8424. doi: 10.1073/pnas.0730515100.

[45] A R Mushegian and E V Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10268–10273, 1996. ISSN 00278424. doi: 10.1073/pnas.93.19.10268.