Universidade Federal do Rio Grande do Sul
UFRGS

Dissertação de Mestrado

Modelo Híbrido de Rede de Associação Proteica.

Ricardo Melo Ferreira

Dissertação apresentada ao Programa de Pós-Graduação em Física da Universidade Federal do Rio Grande do Sul, sob orientação do Profa. Rita Maria Cunha de Almeida e coorientação de Leonardo Gregory Brunnet, como preenchimento parcial dos pré-requisitos para a obtenção do título de Mestre em Física.

Porto Alegre, Março de 2011

Resumo

Dado o grande número de proteínas presentes em um organismo, as associações funcionais entre elas são estudadas através da teoria de redes, com intuito de verificar suas propriedades estatísticas. Diversos modelos de crescimento têm sido propostos para representar as redes de associação proteica. Estes buscam principalmente reproduzir uma distribuição de grau na forma de lei de potência. Entretando, a distribuição de grau das redes de associação proteica não é uma lei de potência pura e outras medidas da rede, como alta clusterização e assortatividade, também não são adequadamente reproduzidas por estes modelos. Neste trabalho procuramos reproduzir as propriedades topológicas das redes de associação proteica. Propomos um modelo baseado em premissas biológicas que considera como fundamental a topologia local da rede para determinar a regra de crescimento. Este modelo reproduz as principais medidas das redes de associação proteica, e a investigação da proposta topológica local pode contribuir para a compreensão biológica dos mecanismos de crescimento do genoma.

Abstract

Given the large number of proteins in a living being, the functional associations between them are studied as networks, in order to verify its statistical properties. Many models of growing networks have been proposed to represent the protein association networks. These models try to reproduce a power law degree distribution. However, the degree distribution of the protein association networks are not a pure power law, and other measures, such as high clustering coefficient and assortativity, are not correctly reproduced by these models. In this work we try to reproduce the topological properties of protein association networks. We propose a biologicaly based model which considers the local network topology as a fundamental ingredient in determining the network growth rule. This model reproduces the essential measures of the protein association networks and the investigation of the local topology proposition may contribute to the biological understanding of the genome growth mechanisms.

Agradecimentos

Gostaria de agradecer primeiramente aos professores Rita de Almeida e Leonardo Brunnet, que além de excelentes orientadores são exemplos como profissionais e pessoas. Obrigado por tudo o que me ensinaram nos últimos anos, que vai muito além do que está contido nestas páginas. Agradeço também aos colegas de grupo pelas reuniões e discussões, sempre muito produtivas. Em especial ao Rodrigo, por esclarecer diversas dúvidas relacionadas à biologia e ter muitas vezes apontado direções à serem exploradas, e ao José Luiz, por ajudar em detalhes técnicos e estar sempre disposto a quebrar um galho. Este trabalho deve-se, também, à excelente formação que tive, graças aos maravilhosos professores dos Institutos de Física e Matemática desta universidade, que dispõem aos alunos todo o conhecimento que podem. A estes meu muito obrigado, sinto apenas não ter aproveitado tanto quanto deveria algumas excelentes disciplinas. Parte desta formação devo aos colegas e amigos que fiz durante os anos de graduação e mestrado. Aos colegas da sala M208 pelas discussões, pelas divertidíssimas tardes, e pelo eventual cafezinho. Aos amigos do "gente esquisita" e aos colegas da infelizmente extinta sala dos bolsistas pelo estudo em grupo, pelas risadas, pelos passeios, pela amizade. Por diversas questões, principalmente a distância e a restrição de tempo, infelizmente acabei nestes anos me afastando de pessoas que muito contribuíram para minha formação pessoal e, portanto, por chegar até aqui. Aos colegas por um curto prazo de tempo do Isshinkan Dojo e ao Sensei Furusho, é impressionante o aprimoramento pessoal que me proporcionaram em tão pouco tempo. Aos grandes amigos, Henrique, Masa, Luiz, Tiago, Josi, Ana, Jéssica e Lisnara. Pelas partidas de RPG e basquete, pelas conversas, pelos estudos, pelas aulas, pela amizade incondicional que me recebe mesmo depois de longos períodos afastado. Posso me considerar um cara de sorte por ter quatro famílias,

e quero agradecer à todas elas. Às famílias do meu pai e da minha mãe, avós, tios e primos. Aprecio muito a companhia de todos, pena que esta não é muito frequente. À minha família emprestada de Passo Fundo, pessoas maravilhosas e divertidíssimas, que sempre se preocuparam comigo. À família da minha namorada, que me acolheu como um filho, obrigado pelo imenso carinho. Agradeço a uma grande amiga que tenho a muito tempo, minha irmã Renata, que muitas vezes me ensinou e me educou, quase sempre através do exemplo. Obrigado por todo o apoio que já me destes, pricipalmente quando mudei-me para Porto Alegre. Ao meu pai, exemplo de pessoa, de honra, de respeito. Eterno incentivador e apoiador, que às vezes se orgulha tanto de minhas conquistas que às torna maiores do que são. A minha mãe, que é uma das pricipais responsáveis pela minha escolha profisional, incentivadora constante de todas as minhas curiosidades, pela educação moral que me deu, por sempre apoiar minhas escolhas, mesmo que não as entendesse, pelo enorme amor e carinho que me deu enquanto esteve ao meu lado. Finalmente, agradeço à minha namorada Débora, meu Raio de Sol, que me apoiou diretamente muito antes de eu escolher a física. Uma companheira e amiga, que com exemplo de dedicação e esforço me faz continuar muito além do ponto onde teria desistido sozinho, a contribuição que deu para a minha formação, profissional e pessoal, é bem maior do que é capaz de perceber. Meu amor, minha amiga, minha colega, minha companheira, meu exemplo, obrigado por voar ao meu lado.

Sumário

Capítulo 1

Introdução

Vivemos em um mundo altamente conectado. Rapidamente nos surge à mente a rede mundial de computadores e a sua capacidade de nos ligar a servidores do outro lado do globo. Mas a internet também trouxe à vida quotidiana o conceito de um outro tipo de rede, o de redes sociais. Na verdade, ao analisarmos cuidadosamente, podemos construir redes em diversas áreas, como em conexões de transporte aéreo, ou redes de transmissão de energia. De fato, qualquer sistema pode ser analisado através da teoria de redes, basta que sejamos capazes de descrever seus elementos e interações por um conjunto de nós e as ligações entre eles, e temos a disposição poderosas ferramentas teóricas para analisar as propriedades estatísticas provenientes das suas interações.

Com base neste princípio, as associações entre genes e proteínas passaram a ser estudadas como uma rede, buscando compreender como a estrutura dessa rede reflete nas funções celulares. Para compreender melhor as redes de associação proteica foram propostos modelos que reproduzissem ou explicassem a sua topologia. Como os organismos estão em constante evolução, nada mais natural que estes modelos construíssem a rede adicionando nós de forma gradual. O modelo Barabási-Albert foi proposto para explicar uma distribuição de grau que seguisse uma lei de potência [1]. A base deste modelo é o princípio de adesão preferencial no qual um novo nó na rede tem uma probabilidade maior de ligar-se aos nós mais conectados. O modelo de Duplicação-Divergência foi proposto para reproduzir uma lei de potência através de um princípio mais adequado biológicamente, uma vez que as redes da associação proteica crescem principal-

mente através de duplicação [2]. Um nó é escolhido com probabilidade uniforme, e um novo nó é adicionado com todas as ligações do anterior, algumas ligações são então removidas, de forma que cada nó do par duplicado tenha alguns vizinhos distintos.

Aqui é inevitável que invertamos a associação que fizemos entre sistemas e redes, e nos perguntemos o que estes modelos representam para o sistema. Uma vez que temos uma regra de adição de nós na rede, o que esta regra representa para os genes e proteínas do organismo? Se encontrarmos um modelo que reproduza corretamente as redes de associação proteicas, teremos uma indicação da evolução dos próprios genes do organismo. Uma vez que estamos tratando com um sistema extremamente complexo, não é natural que exista uma regra universal de adição de nós. Os genes poderiam evoluir de formas diferentes dependendo da função que desempenhassem, ou então, que esta evolução dependesse do organismo em questão, ou do número de genes.

Dado que a regra de crescimento adequada pode servir de refêrencia para a investigação dos mecanismos de crescimento das redes de associação proteica, consideramos fundamental um modelo capaz de reproduzir corretamente estas redes. Os modelos existentes até o momento não obtêm algumas propriedades que apresentaremos neste trabalho, como uma alta distribuição de grau para grau alto, ou a estrutura modular obtida pelo ordenamento, como também não consideram a relação entre topologia e crescimento da rede.

Neste trabalho, buscaremos um modelo de crescimento de redes capaz de reproduzir as propriedades das redes de associação proteica, baseando a nossa regra de adição de nós nos mecanismos biológicos subjacentes à evolução dessas redes.

Capítulo 2

Estado da Arte

Neste capítulo apresentaremos algumas definições necessárias para o estudo de redes de associação proteica. Começaremos com um breve resumo das características biológicas subjacentes a estas redes, bem como apresentando o conceito de redes de associação proteica. Faremos então uma apresentação de redes, das principais medidas utilizadas para definir redes de associação proteica e encerraremos com os principais modelos para representá-las.

2.1 Evolução e Genética

O objetivo da física pode ser definido como o estudo do universo e da natureza. Devido a esse objeto de estudo, ela frequentemente oferece a oportunidade de confrontar-nos com fenômenos que capturam a nossa admiração. Podemos calcular a trajetória de corpos que orbitam a velocidades alucinantes ao redor de estrelas, entender como um inseto é capaz de caminhar colado ao teto ou nos questionar sobre o comportamento dual de entidades subatômicas, apenas para citar algumas destas maravilhas.

Dos frutos da natureza, um dos mais impressionantes é a diversidade da vida que cobre o nosso planeta. Inevitavelmente, a curiosa mente humana começou a formular perguntas a respeito da origem desta diversidade, que tentaram responder filosófica, religiosa e cientificamente. Dois passos foram fundamentais para a construção de uma explicação científica para a abundância de espécies vivas. Em 1859, Charles Darwin publicou o livro "A Origem das Espécies" [3], no qual

afirmava que na disputa por recursos, o animal mais apto sobreviveria e que qualquer modificação, se fosse útil, seria preservada. Em 1865, Gregor Mendel estudava a hibridização de ervilhas e como as características de uma ervilha são passadas aos seus descendentes [4]. Estas são duas sementes de uma explicação científica para a grande variedade de espécies. As características de um indivíduo são passadas aos seus descendentes através de seus genes, e uma pequena mutação nos genes passados a algum descendente pode resultar em uma característica que será preservada. Se dois grupos de uma mesma espécie se mantiverem separados, o número de mutações acumuladas ao longo de várias gerações pode ser tal que os grupos se tornem espécies diferentes.

2.1.1 ADN

O grupo de genes de cada indivíduo é chamado genoma e está em seu ADN, ácido desoxirribonucleico, que possui quatro diferentes bases nitrogenadas, adenina, timina, citosina e guanina. Ele está armazenado nas células no formato de uma longa molécula chamada cromosomo. O cromossomo é composto de duas fitas de ADN entrelaçadas, que são basicamente sequências de bases nitrogenadas. Cada base em uma das fitas se liga au seu par na fita complementar, adenina se liga com timina e citosina com guanina, de forma que a sequência de bases de uma fita é completamente determinada pela complementar. Assim temos um mecanismo que permite a replicação do código genético, desenrolando as duas fitas e utilizando cada uma para produzir uma cópia da complementar. Genes são os grupos de bases do ADN que codificam a informação genética. Há também regiões de bases de ADN intercaladas a esses grupos que são não codificantes, estas podem ter funções regulatórias ou estruturais [5]. A principal função dos genes é codificar a construção das proteínas utilizadas nas reações químicas celulares. A transcrição desse código é feito através do ARN, ácido ribonucleico, que direciona a construção do proteoma, o conjunto de proteínas codificadas pelo genoma. O conjunto de moléculas de ARN de uma célula é chamado de transcriptoma. Para codificar as proteínas, as bases nitrogenadas são lidas em grupos de três, chamados códons que codificam aminoácidos utilizados para construir as proteínas. Uma vez construídas as proteínas dobram-se, formando sítios de ligação, que podem se ligar a outras proteínas ou a determinados genes. Essa ligação se dá por interações físico-químicas entre

dos sítios que se complementam, como num par chave-fechadura. Estas proteínas podem ter as mais variadas funções; podem interagir com outras proteínas, formar estruturas celulares, atuar como enzimas que reparam o ADN e até mesmo desempenhar atividades regulatórias durante a transcrição do genoma.

2.1.2 Mutações

O ADN de uma célula está constantemente sofrendo alterações em sua sequência de bases nitrogenadas, chamadas mutações. Algumas dessas alterações são erros espontâneos durante a replicação do ADN, necessária para divisão celular, e outras são provocadas por agentes mutagênicos externos ao ADN, como radiação ultra-violeta ou substâncias químicas que induzam um erro na replicação. Mutações em regiões não codificantes são silenciosas, ou seja, não produzem alteração no proteoma, mas uma mutação em um gene pode ter diferentes resultados [5].

No caso de mutações pontuais, em que uma única base nitrogenada é trocada por outra, quatro combinações são possíveis: podemos ter uma mutação sinônima, em que o códon resultante codifica o mesmo aminoácido; mutação em que apenas um aminoácido da proteína é alterado, geralmente sem consequências drásticas para a função da proteína; pode-se transformar um códon que especifica um aminoácido em um códon de parada, resultando assim numa proteína mais curta, usualmente não funcional; a mutação pode tranformar um códon de parada em outro que codifica um aminoácido, resultando em proteínas mais longas, que podem ter problemas em se dobrar e perder parte da atividade. Deleções e inserções de bases também podem ter resultados diferentes. Caso a alteração no número de bases seja múltiplo de três, pode resultar apenas em uma troca em um grupo de aminoácidos, e a proteína pode não sofrer uma grande alteração. Caso contrário, todos os códons depois do ponto de inserção ou deleção serão codificados de forma incorreta, resultando em uma proteína completamente diferente.

Uma mutação que altere significantemente uma proteína pode ter diferentes efeitos em um organismo. No caso de um ser unicelular, este pode se tornar dependente de uma determinada substância que é produzida por uma célula sem mutação, ou pode adquirir resistência a algum tóxico. Quando essa mutação não resultar na morte do organismo, a propriedade adquirida passa

a integrar o genoma de seus decendentes. Em seres pluricelulares, a mutação só é passada às próximas gerações caso ocorra em um gameta. Em uma célula somática, mesmo que a mutação resulte em morte celular, dificilmente será perceptível no organismo, a não ser que resulte em atividade cancerosa [6]. No caso de mutações em gametas, o organismo pode deixar de ser viável, perder ou ganhar funções. Neste último caso, a alteração estará presente em toda a descendência do organismo.

2.1.3 Mutações e a Evolução do Genoma

A mutação nos dá um mecanismo para a evolução dos genes, mas sozinha não é capaz de explicar completamente a evolução do genoma, principalmente a diferença no número de genes em diferentes organismos. O acréscimo de genes se dá por diversos mecanismos [7], dos quais podemos citar como mais importantes o surgimento de genes *de novo*, a retroposição e a duplicação de genes [2].

O surgimento de genes *de novo* ocorre quando uma região não codificante do genoma sofre uma mutação e passa a codificar um gene [8, 9]. Um gene surgido por esse fenômeno pode ser identificado comparando a sequência de bases deste gene com regiões não codificantes de organismos da mesma família. Uma vez que um gene é tranScrito em ARN, ele pode ser novamente codificado em ADN e reinserido no genoma, criado uma cópia do gene original. Esse mecanismo de surgimento de genes é chamado retroposição [10]. Os genes surgidos por retroposição tem como característica a inexistência de introns, pequenos segmentos não codificantes dentro do gene, que não são transcritos em ARN. Apesar de a retroposição criar uma cópia do gene, o mais comum é a duplicação de genes: o surgimento de cópias idênticas da sequência de bases nitrogenadas, causado, por exemplo, por um erro durante a replicação do genoma. A duplicação de genes é fundamental para a evolução do genoma, pois oferece matéria prima para mutação. Uma mutação que seria letal para o organismo, por exemplo, agora pode ocorrer em uma das cópias uma vez que a outra cópia manteria a função do gene original. Isso permite que cada um dos genes assuma parte da função exercida pelo gene ancestral, e adquira também novas funções [11–15]. Outro mecanismo de aquisição de genes, de menor impacto no genoma, é a

transferência horizontal de genes, no qual os genes são copiados de outra espécie. Apesar de ocorrer com maior frequência em organismos procariotos [16, 17], com a disponibilidade cada vez maior de dados a respeito do genoma dos organismos foram encontrados diversos casos de transferência horizontal em eucariotos [18].

2.2 Redes

Uma rede pode ser definida de forma abstrata como um conjunto de nós, e um conjunto de ligações entre esses nós [19]. De forma genérica, uma rede pode ter mais de um tipo de nó, e as ligações podem ter uma direção ou um peso atribuído. Em uma rede de citações bibliográficas, por exemplo, as ligações são direcionadas de um artigo para as suas referências, já uma rede de predadores e presas pode ter mais de um tipo de nó definido, como plantas, herbívoros e carnívoros. Esses dois exemplos, além de ilustrar algumas características, ilustram a variedade de assuntos que podem ser abordados com as ferramentas da teoria de redes. Com efeito, estas ferramentas tem sido aplicadas a inúmeros fenômenos distintos, como propagação de epidemias, colaboração de atores em filmes, linguística, redes de distribuição elétrica, internet e redes de associação proteica [20]. Utilizar uma única abordagem para problemas tão variados somente é possível porque os elementos e as interações são representados por nós e ligações sem distinção da natureza daqueles. Os objetivos, então, da teoria de redes, focam-se em medidas estatísticas, modelos de formação, e estudo de sistemas dinâmicos nas redes.

2.2.1 Medidas

Grau

Apresentaremos algumas medidas utilizadas no estudo de redes [21]. Ao longo desse trabalho consideraremos redes com ligações sem peso, não direcionadas e com um único tipo de nó. A generalização destas medidas para casos específicos é simples e pode ser encontrada em Costa *et al.* [22]. Uma rede com n nós pode ser completamente definida por uma matriz quadrada A_{nn}, chamada matriz de associação, representativa das ligações entre cada par de nós da rede: $A_{i,j} = 1$

indica a existência de uma ligação entre os nós i e j; $A_{i,j} = 0$ a sua inexistência. Como as ligações não são direcionadas a matriz de associação é simétrica. A primeira característica de um nó i é o seu grau k_i, que representa o seu número de ligações, e é dado por

$$k_i = \sum_{j=1}^{n} A_{i,j} \ , \tag{2.1}$$

podemos também definir o grau médio da rede como

$$\langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i \ , \tag{2.2}$$

que é a razão entre o número de ligações e o número de nós.

Distribuição de Graus

As ligações entre os nós formam a estrutura da rede, mas para uma rede com um grande número de nós não podemos analisar o grau de cada nó independentemente, e o grau médio não nos fornece muita informação sobre a estrutura da rede. Definimos, então, $N(k)$, a fração de nós com grau k, de forma que

$$N(k) = \frac{1}{n} \sum_{i=1}^{n} \delta_{k_i,k} \ . \tag{2.3}$$

Esta fração pode também ser interpretada como a probabilidade de um nó escolhido aleatoriamente ter grau k. Para podermos comparar facilmente redes de tamanhos diferentes, apresentaremos as medidas reescalando o grau por $1/k_{mx}$, onde k_{mx} é o maior grau presente na rede. Para mantermos a normalização, definiremos então a distribuição de grau como

$$P(k) = k_{mx} N(k) \ . \tag{2.4}$$

A distribuição de grau pode ser utilizada para determinar a estrutura da rede. Por exemplo, uma rede aleatória possui uma distribuição de grau dada pela distribuição de Poisson, e redes com uma distribuição exponencial de grau, na forma $P(k) \propto k^{-\gamma}$ ($\gamma > 0$), são as chamadas redes livres de escala(*scale-free*).

Correlações entre Graus

Outra forma de caracterizar estatisticamente a rede é procurar correlações entre os graus. A forma mais imediata de fazê-lo é determinando o grau médio dos vizinhos dos nós de grau k. Fazemos isso primeiramente calculando o grau médio dos vizinhos de cada nó i, dado por

$$k_{nn,i} = \frac{1}{k_i} \sum_{j=1}^{n} A_{i,j} k_j \ . \tag{2.5}$$

Como $A_{i,j} = 0$ caso os nós i e j não estejam ligados, a soma é feita somente sobre os vizinhos do nó i. Determinamos então grau médio dos vizinhos dos nós de grau k pela relação

$$k_{nn}(k) = \frac{1}{N_k} \sum_{i=1}^{n} \delta_{k,k_i} k_{nn,i} \ , \tag{2.6}$$

onde N_k é o número de nós de grau k, e a soma é feita sobre os nós i que satisfaçam $k_i = k$. Este valor está diretamente relacionado com a probabilidade condicional $P(k'|k)$ de que uma ligação de um nó de grau k leve a um grau k'

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \ . \tag{2.7}$$

Se os graus forem completamente descorrelacionados, a probabilidade condicional será função apenas de k', e portanto $k_{nn}(k)$ é uma constante. Dois comportamentos são possíveis para esta medida: caso $k_{nn}(k)$ aumente com k, os nós de maior grau tendem a se ligar com nós de grau também alto, a rede é chamada assortativa; caso diminua, os nós de maior grau ligam-se com nós de baixo grau, e a rede é dissortativa.

Ciclos

Outra medida estatística em redes é a identifição de ciclos. Um ciclo é um caminho fechado, percorrido através das ligações entre os nós. Seu tamanho é medido pelo número (mínimo) de ligações necessárias para se voltar a um nó de partida sem repetir passagens sobre outro nó qualquer. Em rede sociais, por exemplo, é frequente a presença de ciclos de tamanho três e é fácil

perceber que isto ocorre. Se escolhermos dois amigos de determinado indivíduo, há uma grande probabilidade de eles serem também amigos entre si. Uma forma de medir a presença destes ciclos de tamanho três é a *clusterização*. O número total de ciclos de grau três N_3 que poderiam passar por um nó i é a combinação de seus vizinhos tomados dois a dois, $N_3 = k_i(k_i - 1)/2$. Já o número de ciclos de tamanho três que passam por este nó é o número de ligações entre os seus vizinhos, l_{ni}. A clusterização C_i do nó é então dada por

$$C_i = \frac{2l_{ni}}{k_i(k_i - 1)} \ . \tag{2.8}$$

Mas assim como o grau, esta é uma medida local, e a clusterização individual de cada nó não é útil para caracterizar grandes redes. Podemos, então, definir a clusterização média da rede

$$\langle C \rangle = \frac{1}{n} \sum_{i=1}^{n} C_i \ , \tag{2.9}$$

mas novamente perdemos informação sobre a estrutura da rede. Utilizaremos então a clusterização média por grau, dada por

$$\langle C \rangle_k = \frac{1}{N_k} \sum_{i=1}^{n} \delta_{k,k_i} C_i \ . \tag{2.10}$$

Ordenamento por Função Custo

Para caracterizar a rede também utilizaremos o algoritmo de ordenamento introduzido por Rybarczyk-Filho *et al.* [23], que permite vizualizar padrões estruturados diretamente na matriz de associação. O objetivo do algoritimo é reordenar as linhas e colunas dessa matriz minimizando uma função custo dada por

$$E = \sum_{i,j=1}^{n} |i - j| \left(|A_{i,j} - A_{i+1,j}| + |A_{i,j} - A_{i-1,j}| + |A_{i,j} - A_{i,j+1}| + |A_{i,j} - A_{i,j-1}| \right) \ , \tag{2.11}$$

Os termos da forma $|A_{i,j} - A_{k,l}|$ representam o módulo da diferença entre os dois elementos da matriz, enquanto $|i - j|$ é proporcional à distância do ponto até a diagonal da matriz. A idéia desse algoritmo é penalizar interfaces entre 1 e 0 na matriz, e penalizar ainda mais quando esta

interface estiver longe da diagonal. Utilizaremos uma função custo ligeiramente modificada

$$E = \sum_{i,j=1}^{n} |i-j|^{\alpha} \left(|A_{i,j} - A_{i+1,j}| + |A_{i,j} - A_{i-1,j}| + |A_{i,j} - A_{i,j+1}| + |A_{i,j} - A_{i,j-1}| \right) \ , \qquad (2.12)$$

onde o expoente (neste trabalho usamos $\alpha = 8$) força as ligações a ficarem mais próximas da diagonal. O algoritmo de ordenamento sorteia dois nós e os troca de posição na matriz de associação. Aceitamos a troca utilizando *annealing simulado* [24]. Caso a função custo diminua, a troca é aceita. Caso aumente, é aceita com uma probabilidade dada por $e^{\delta E/T}$, onde δE é a variação do custo e T uma temperatura virtual. Começamos com uma temperatura de 10^9, e diminuimos por 20% a cada 100 passos de MonteCarlo, para redes com $n \approx 10^4$ nós.

As medidas utilizadas e o algoritmo de ordenamento ficarão mais claros à medida que apresentarmos os modelos de redes e as redes de associação proteicas. Para permitir que várias redes possam ser analisadas em conjunto, reescalamos o grau por $1/k_{max}$ onde k_{max} é o maior grau presente na rede.

2.3 Modelos de redes de associação proteica

As medidas definidas acima são macroscópicas, ou seja, dão informações globais sobre a rede, e não sobre a a estrutura de ligações de cada nó individualmente. Então, uma vez que tenhamos caracterizado as redes reais através das medidas apresentadas acima, o próximo passo é buscar modelos de redes que também as reproduzam, pois este modelo poderá indicar a estrutura local de cada nó. Diversos modelos têm sido propostos para redes de associação proteica [25–28]. Em sua maioria estes modelos são de crescimento de redes nos quais nós são adicionados na rede seguindo uma determinada regra, uma vez que o genoma, e portanto o número de proteínas de um organismo, cresce gradualmente. Dos modelos existentes apresentaremos os de Barabási-Albert [1, 29] e o de Duplicação-Divergência [30, 31]. No capítulo 5 compararemos seus resultados com medidas de redes de associação proteica, e no capítulo 4 proporemos um modelo de crescimento de redes baseado nesses dois mecanismos.

2.3.1 Modelo de Barabási-Albert

O modelo de Barabási-Albert foi uma primeira proposta para explicar diversas redes, como redes de associação proteica, internet, colaboração de atores, entre outras, que tenham em comum o fato de serem redes em constante crescimento e possuirem distribuição de grau de lei de potência. Este é um modelo de crescimento no qual a rede cresce segundo a seguinte regra: dada uma pequena rede inicial com m_0 nós, cada novo nó i é adicionado com m ligações. O outro nó j dessas ligações é escolhido com probabilidade proporcional a k_j. Assim o nó com maior grau tem maior probabilidade de receber uma nova ligação. Esse mecanismo é chamado de adesão preferencial, e é de fácil justificativa em redes como a de colaboração de atores em filmes, uma vez que um novo ator tem uma grande probabilidade de ser escalado para um filme com outro ator mais conhecido. Este modelo produz uma rede livre de escala com uma distribuição de grau dada por $P(k) = 2m^2/k^3$ com expoente $\gamma = 3$. Na figura 2.1 temos as medidas do modelo de Barabási-Albert para $m = 3$ e uma rede inicial de $m_0 = 5$ nós ligados em anel.

Na figura **2.1 a** temos a distribuição de grau característica de redes livres de escala. Na figura **2.1 b** temos uma clusterização média por grau baixa, em relação ao velor máximo um. Isto ocorre pois, uma vez que os vizinhos dos novos nós são escolhidos através de adesão preferencial, há uma pequena probabilidade de que o novo nó se ligue a dois vizinhos, formando um triângulo. A figura **2.1 c** mostra o crescimento da rede, que tem um grau médio constante, uma vez que cada nó é adicionado com três ligações. Finalmente a figura **2.1 d** mostra que a rede obtida é dissortativa. Como cada novo nó tende a ligar-se com os nós de maior grau, o grau médio dos seus vizinhos é alto. Com os nós de maior grau ocorre o oposto, como recebem muitos vizinhos novos, o grau médio de seus vizinhos é baixo.

Na figura 2.2 temos o resultado do algoritmo de ordenamento aplicado à matriz de associação da rede, gerando uma nova matriz de associação. Essa matriz será utilizada para comparação de estruturas com as obtidas para os organismos.

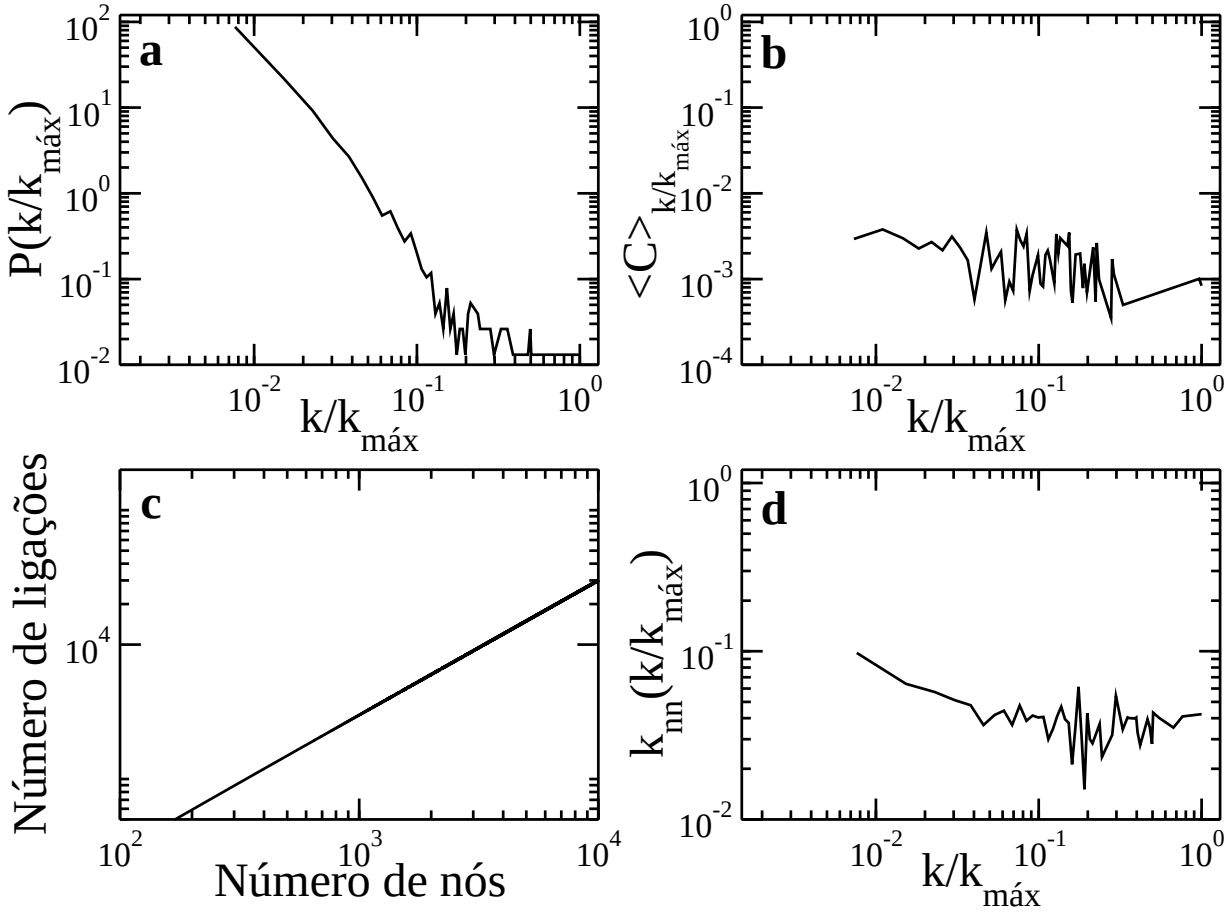

Figura 2.1: Caracterização de rede obtida pelo modelo de Barabási-Albert com $m = 3$ e $m_0 = 5$

2.3.2 Modelo de Duplicação-Divergência

Este modelo foi proposto por Vázquez *et al.* [30, 31] como uma alternativa ao modelo Barabási-Albert para a geração de redes de associação proteica. A regra de crescimento da rede foi inspirada no processo de duplicação e complementação funcional do genes. O processo de adição dos nós ocorre em duas etapas:

- Duplicação: um nó i é escolhido aleatoriamente e um nó i' é adicionado à rede com todas as ligações iguais às do nó i. Uma ligação entre os dois nós é adicionada com probabilidade p.

- Divergência: para cada nó j ligado à i e i', escolhemos aleatóriamente uma das duas ligações e eliminamos com probabilidade q.

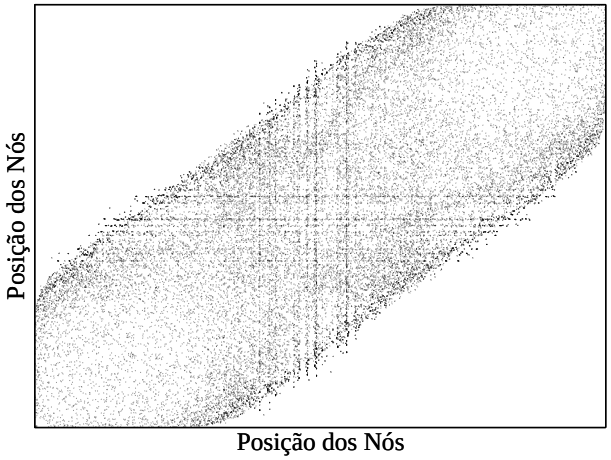

Figura 2.2: Matriz de associação ordenada da rede obtida pelo modelo de Barabási-Albert com $m = 3$ e $m_0 = 5$

A relação destes processos com o mecanismo biológico é simples. Como os nós são idênticos, a diferença entre duas proteínas é representada simplesmente pela sua vizinhança. Como os genes recém duplicados são idênticos, suas proteínas são idênticas, e portanto são representadas por vizinhanças idênticas. A divergência leva em conta o papel da mutação na formação da rede, assumindo que os genes recém duplicados possam sofrer mais eventos de mutação. Cada vizinho do par de nós duplicados vai se manter ligado a um dos nós, como cada ligação é uma associação funcional, isso significa que as funções exercidas pelo nó original serão mantidas pelo par duplicado.

A figura 2.3 foi obtida utilizando $p = 1$ e $q = 0.8$ em uma rede inicial de cinco nós em anel. Escolhemos $p = 1$ por simplicidade, a fim de evitar redes bipartidas, isto é, que tenham duas ou mais componentes sem ligação entre si. Na figura **2.2 a** vemos que o modelo de Duplicação-Divergência leva também a uma distribuição de grau em forma de lei de potência, apesar de não considerar diretamente a adesão preferencial. Isso se deve ao fato que os nós de maior grau tem uma probabilidade maior de terem um vizinho duplicado e, portanto, de receber uma nova ligação. Vemos também em **2.2 b** que a rede obtida tem uma clusterização média por grau mais alta que a rede de Barabási-Albert, mas com uma queda para graus maiores. Em **2.2 c** temos o crescimento desta rede, e em **2.2 d** vemos que a rede é assortativa. Novamente podemos explicar este comportamento. Cada vez que um nó passa por um processo de duplicação, ele ganha um

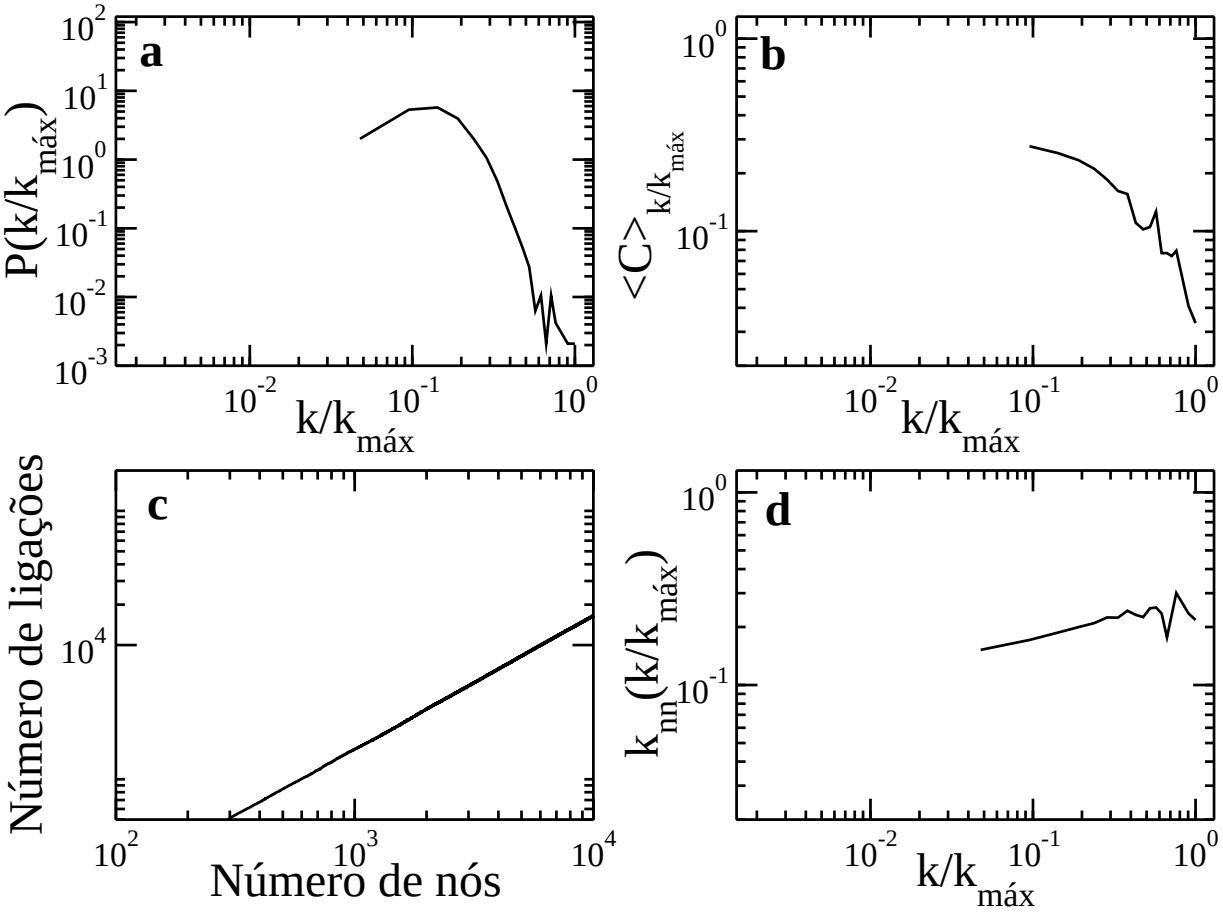

Figura 2.3: Caracterização de rede obtida pelo modelo de Duplicação-Divergência com $p = 1$ e $q = 0.8$

vizinho de grau semelhante ao seu, portanto nós com alto grau ganham vizinhos também de grau alto. No apêndice A encontra-se o resultado para estes dois modelos para diferentes parâmetros.

Na figura 2.4 vemos que a matriz de associação ordenada para a rede do modelo de Duplicação-Divergência possui uma distribuição de ligações aproximadamente uniforme no centro, e nas extremidades as ligações se definem em duas distâncias, uma próxima à diagonal e outra mais distante.

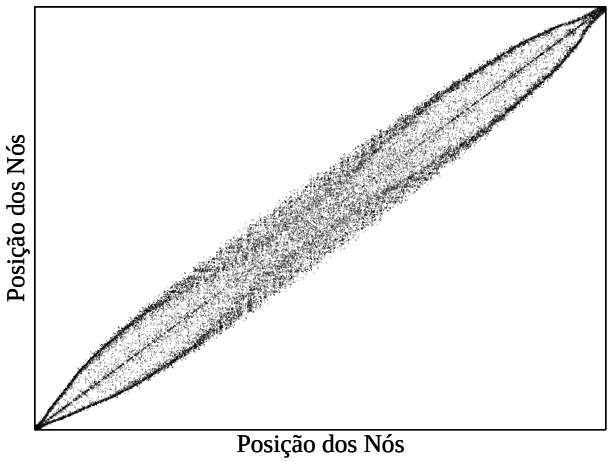

Figura 2.4: Matriz de associação ordenada da rede obtida pelo modelo de Duplicação-Divergência com $p = 1$ e $q = 0.8$

Capítulo 3

Redes de associação proteica

Dado o grande número de genes dos organismos é natural uma abordagem através de teoria de redes. Uma rede de associação proteica é construída atribuindo um nó a cada proteína e uma ligação, caso exista uma associação funcional entre elas. Neste trabalho utilizaremos redes de associação proteica extraídas do banco de dados STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) mantido pelo Laboratório Europeu de Biologia Molecular (*European Laboratory for Molecular Biology - EMBL*) e disponível no sítio *http://string.embl.de/* [32–34]. O STRING utiliza sete métodos de previsão de interação: vizinhança, fusão de genes, coocorrência, coexpressão, experimentos, base de dados e mineração de texto.

- Vizinhança: a conservação de vizinhança em genomas de diferentes organismos é interpretada como uma indicação de associação funcional entre as proteínas codificadas, assim, se um grupo de genes é vizinho há uma probabilidade de interação. A pricipal desvantagem deste método é que este fenômeno ocorre somente em genomas de procariotos.

- Fusão de genes: é prevista uma interação entre duas proteínas de um determinado organismo se elas fazem parte de uma mesma proteína em outro.

- Coocorrência: é baseada na presença de ligação entre as proteínas através das espécies. Se duas proteínas estão ligadas em um grande número de organismos é provável que elas estejam ligadas nos outros em cujos proteomas estejam presentes.

- Coexpressão: este método prediz a existência de uma ligação para proteínas cujos genes apresentem respostas de transcrição em ARN a semelhantes a mudanças de estado ou comportamento celular.

- Experimentos: é uma lista de interações proteicas obtidas experimentalmente, retiradas de bancos de dados.

- Base de dados: é uma lista de associações funcionais obtidas de bancos de dados de rotas metabólicas.

- Mineração de texto: as ligações são inferidas através de artigos. O aparecimento no mesmo texto indica ligação entre as proteínas.

O STRING atribui ainda um índice de confiança S_i para cada método utilizado para prever determinada ligação. Os índices são obtidos testando as previsões com um conjunto confiável de ligações. Os índices individuais são então combinados para se atibuir um índice de confiança para a ligação da seguinte forma

$$S = 1 - \prod_{i=1}^{7}(1 - S_i) \ , \tag{3.1}$$

A seguir caracterizaremos as redes obtidas do STRING, utilizando as medidas apresentadas na seção 2.2. Desconsideramos também o método de predição de mineração de texto, pois ele não representa necessariamente interação entre as proteínas envolvidas.

Na figura 3.1 temos as medidas para as redes de todos os organismos *core* do STRING, para índices de confiança $S = 0.700$, $S = 0.800$ e $S = 0.900$. Os organismos considerados *core* são importantes organismos modelo e para os quais existem dados experimentais. Temos vários pontos para os mesmos graus pois há várias redes (organismos) para cada índice de confança. Além disso, para que todas as redes coincidissem reescalamos o grau por $1/k_{mx}$. A primeira propriedade que podemos observar é que a distribuição de grau não pode ser descrita como simplesmente uma lei de potência, uma vez que a distribuição tem uma elevação para alto grau. Isso significa que há um número considerável de nós com alto grau, ao contrário de uma rede livre de escala, na qual esse número é muito pequeno. Na figura **3.1 b** vemos que as redes de associação proteica são

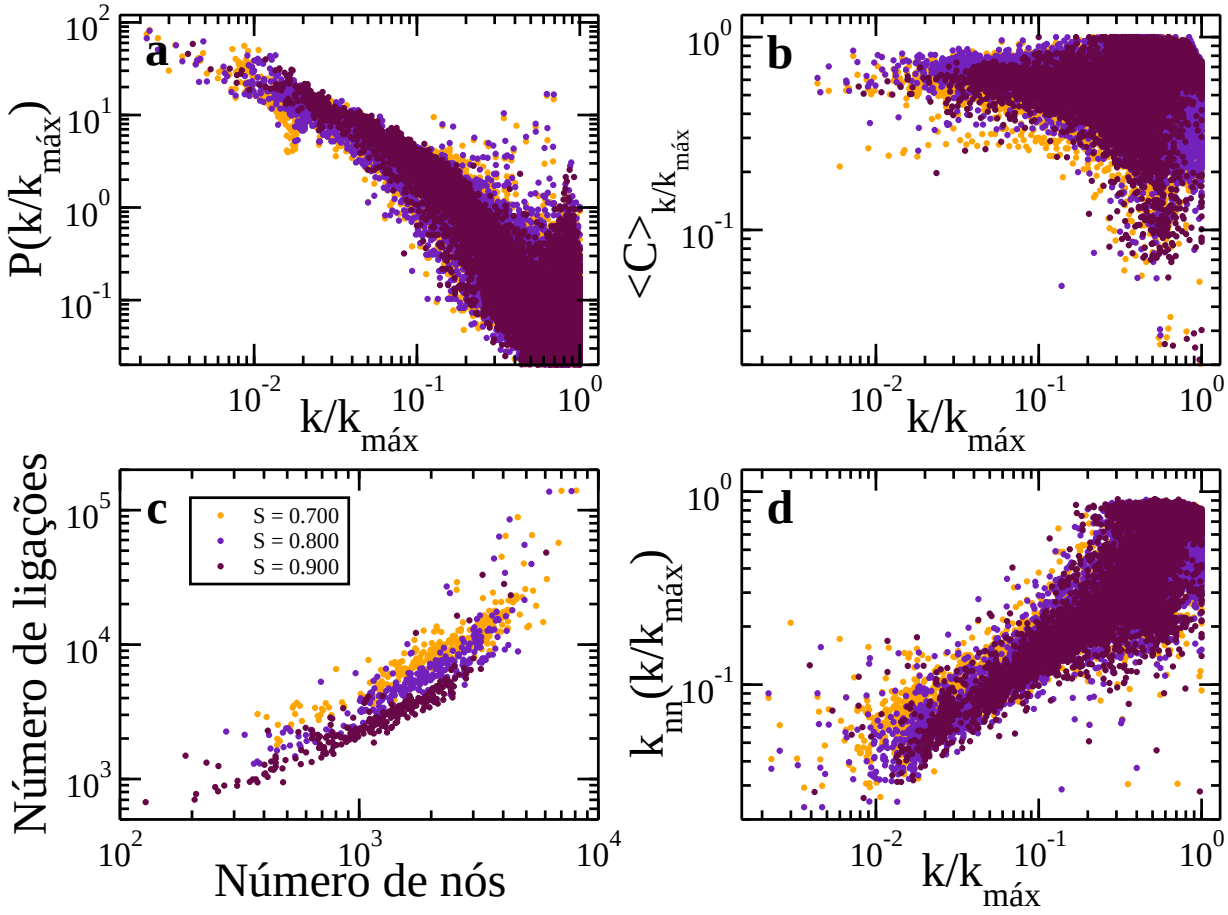

Figura 3.1: Medidas dos organismos classificados como *core* no STRING para três índices de confiança.

altamente clusterizadas, com clusterização média por grau para os graus mais altos próximas de um em alguns casos. Isso significa que temos nós com um grande número de ligações altamente ligados entre si. Em **3.1 c** temos o número de ligações por nó das redes. Note que neste gráfico temos um ponto para cada organismo em cada índice, enquanto que nos demais cada rede dá origem a uma distribuição que são sobrepostas. Na figura **3.1 d** podemos ver que as redes de associação proteica são fortemente assortativas.

Na figura 3.2 temos as mesmas medidas para os organismos *Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens, Mus musculus* e *Escherichia coli* para índice de confiança $S = 0.800$. Estes organismos foram escolhidos pois são muito estudados e, portanto, possuem uma quantidade maior de informação. Observe que na figura **3.2 c** temos, além dos organismos escolhidos, os outros organismos para os três índices de confiança. Nesta

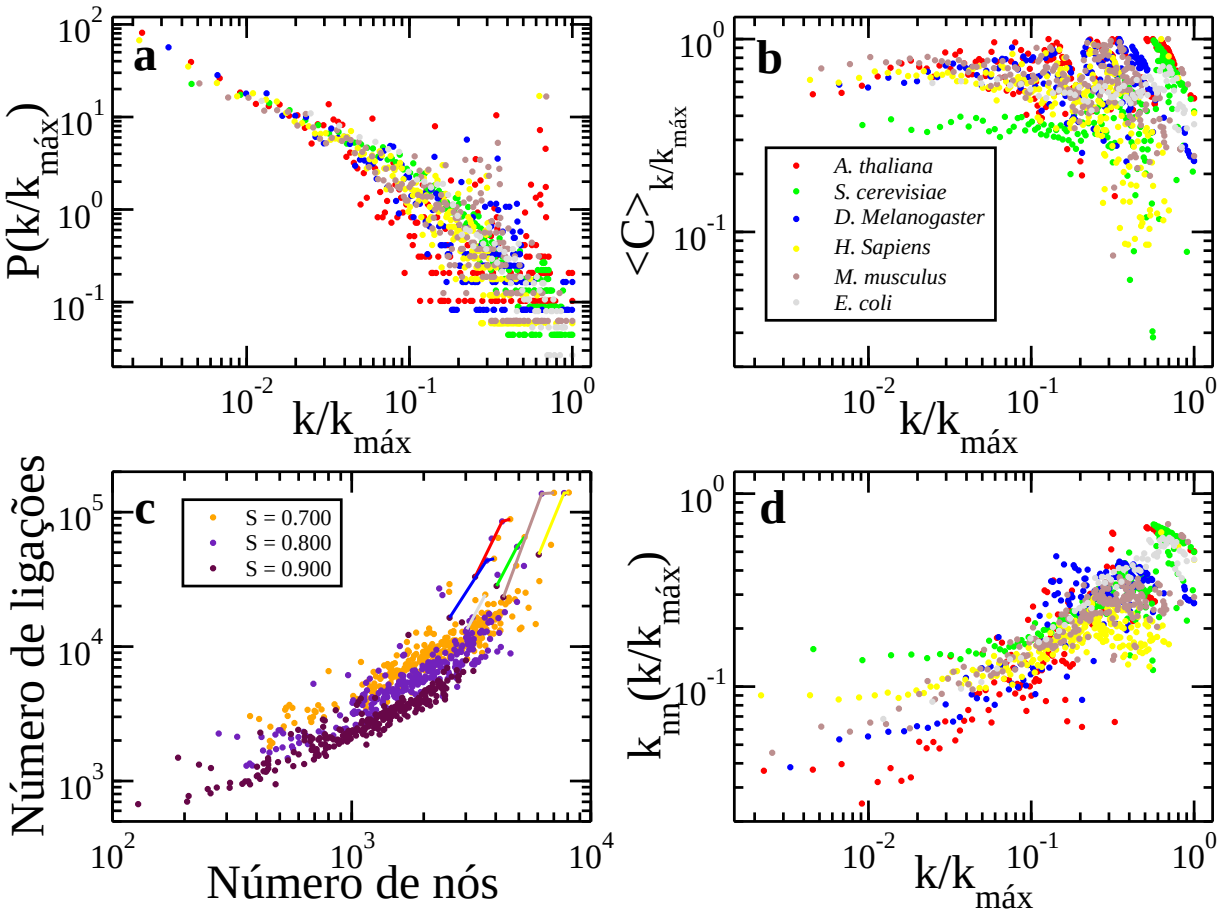

Figura 3.2: Medidas dos organismos *Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens, Mus musculus* e *Escherichia coli* para índice de confiança $S = 0.800$. Na figura **3.2 c** apresenta todos os organísmos para os três índices para comparação.

figura vemos que estes organismos têm um número de ligações acima de outros organismos com número de nós semelhantes. Isto confirma a hipótese que utilizamos para a escolha, pois indica que se conhece um maior número de associações para as proteínas destes organismos. Nas figuras **3.2 a**,**3.2 b** e **3.2 d** vemos que estes organismos representam a estrutura das redes de associação proteica, *i.e.*, uma distribuição de grau acima do esperado para alto grau em uma rede livre de escala, clusterização média alta e fortemente assortativa. Tomaremos estas redes como padrão para apresentar os próximos resultados.

Na figura 3.3 vemos que a matriz de associação ordenada dos organismos possuem algumas propriedades interessantes. As ligações se distribuem em uma forma que lembra uma folha, com a presença de módulos altamente conectados. Em alguns casos, como *Arabidopsis thaliana* e *Homo*

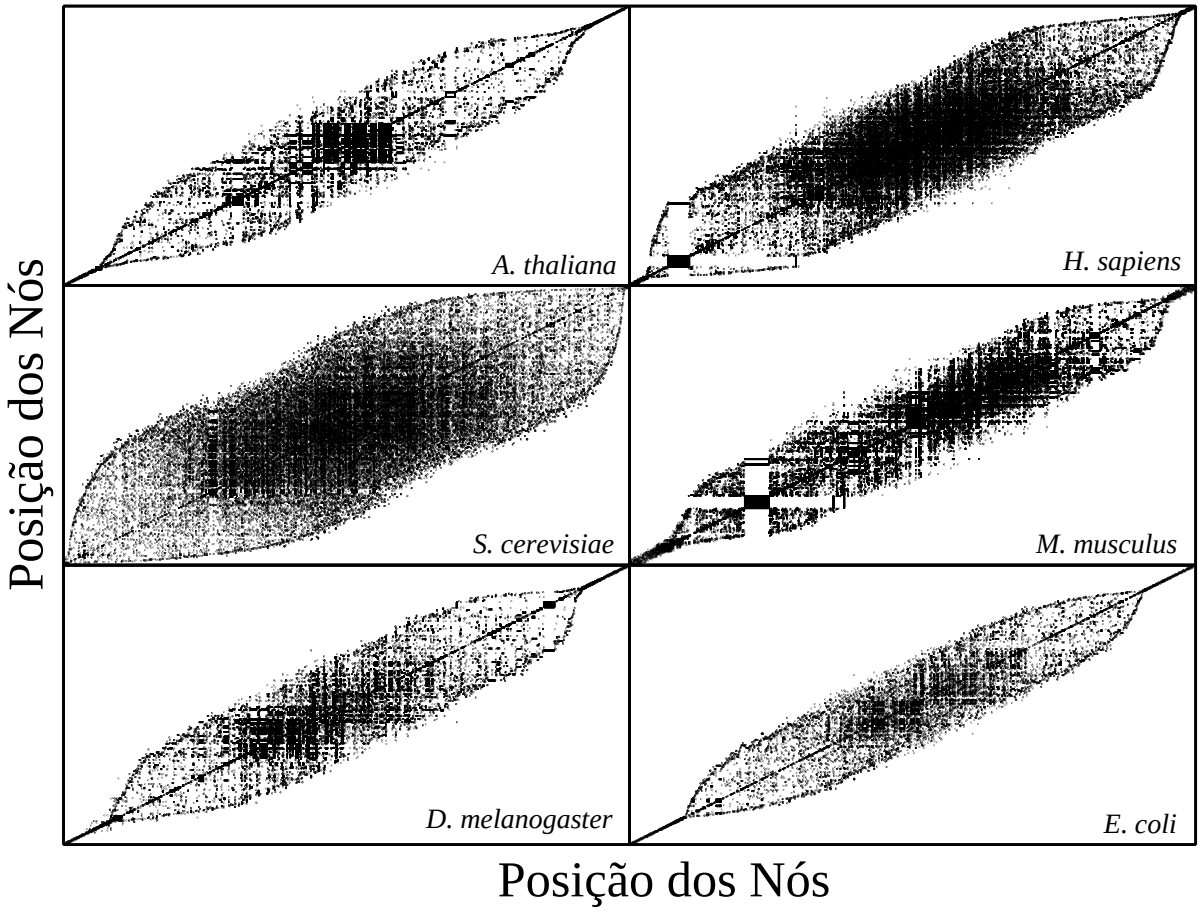

Figura 3.3: Matrizes de associação ordenadas dos organismos *Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens, Mus musculus* e *Escherichia coli.*

sapiens, o algoritmo produz blocos quadrados, o que significa que neste módulo os vizinhos estão próximos no ordenamento e um grande número das ligações destes nós é interna ao módulo.

3.0.3 Comparação entre redes de associação protéica e modelos

Uma vez que temos as redes de associação protéica é natural compará-las diretamente com as redes com 10^4 nós obtidas segundo os modelos de Basabási-Albert e de Duplicação-Divergência. Nas figuras 3.4 e 3.5, temos as redes obtidas com esses modelos sobrepostas às redes dos organismos padrão.

Podemos facilmente notar que as redes obtidas pelos modelos não reproduzem completamente as redes de associação protéica. A rede obtida pelo modelo de Barabási-Albert é dissortativa,

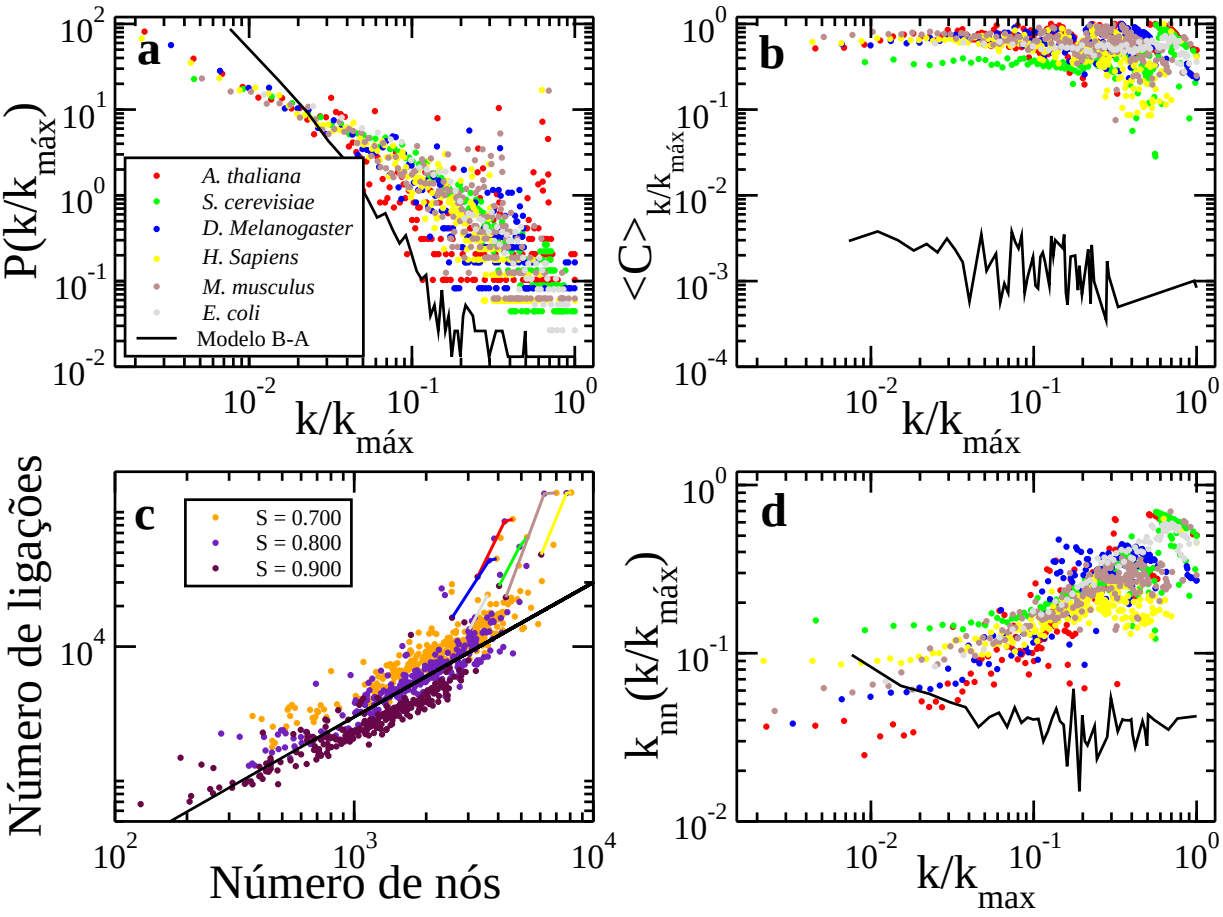

Figura 3.4: Comparação entre as redes do modelo de Barabási-Albert e dos organismos.

e tem clusterização baixa, ao contrário das redes dos organismos. O modelo de Duplicação-Divergência produz uma rede assortativa e com uma clusterização mais alta, mas ainda abaixo da encontrada nas redes de associação proteica. O modelo de Duplicação-Divergência tem ainda um grau máximo k_{max} pequeno, uma vez que é apenas cerca de dez vezes maior que o menor grau da rede, isto é aproximadamente $k_{max} \approx 10$. Por último, podemos notar que ambos possuem uma distribuição de grau com poucos nós para graus mais altos.

Temos então que buscar um modelo de crescimento de rede que seja capaz de reproduzir as redes de associação proteica, com uma distribuição de grau com um aumento no número de nós para alto grau, com clusterização média por grau alta, e fortemente assortativas, caso contrário podemos estar negligenciando fatores importantes para explicar o comportamento dessas redes.

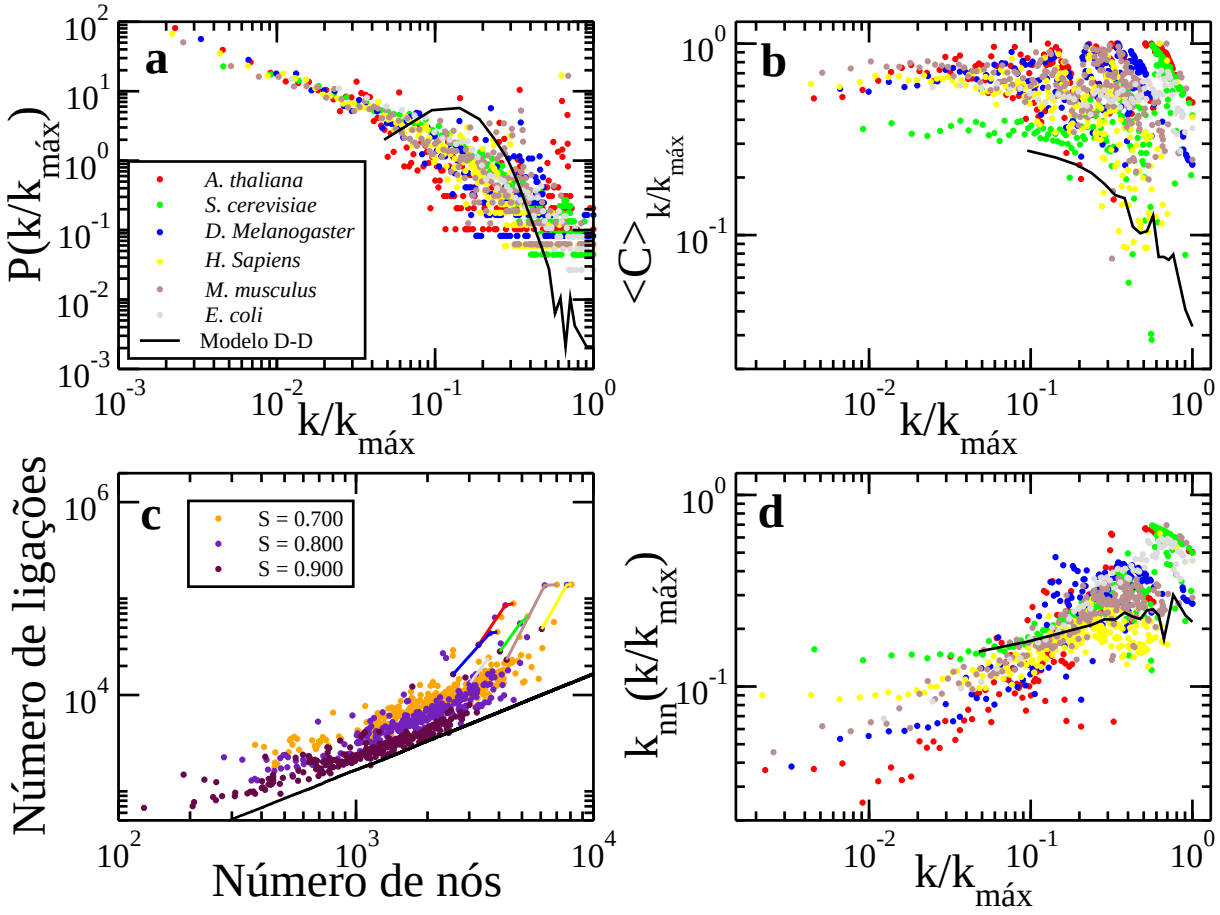

Figura 3.5: Comparação entre as redes do modelo de Duplicação-Divergência e dos organismos.

Capítulo 4

Modelo

Propomos um modelo de crescimento de redes que simula o mecanismo biológico subjacente à evolução das redes de associação proteica, e reproduz as suas propriedades topológicas. Para isso consideramos que os principais mecanismos de surgimento de uma nova proteína na rede podem ser classificados em dois grupos: o surgimento de genes *de novo* e a transferência horizontal, que são o surgimento de um novo gene no genoma sem relação com os demais; e a duplicação e retroposição de genes, que são duas formas de inserir no genoma cópias de genes antigos. Representaremos cada um desses grupos por um mecanismo na construção das redes, e chamaremos esses dois mecanismos de criação e duplicação de nós.

Para estabelecer as ligações para os nós resultantes do processo de criação utilizaremos o princípio da adesão preferencial apresentado na subseção 2.3.1. Podemos supor que as proteínas mais ligadas têm um número maior de sítios de ligação, e que, portanto, uma proteína nova tem uma chance maior de ligar-se a ela. Introduzimos, entretanto, uma pequena modificação no modelo de Barabási-Albert, uma vez que não mantemos fixo o número m de ligações do novo nó. Segundo Newman [19], Price havia proposto um modelo semelhante para redes de associação bibliográfica. Introduzimos um novo nó i na rede da seguinte forma: a probabilidade de um nó j receber uma ligação do nó i é proporcional ao seu grau e dada por

$$pl = \frac{k_i}{\sum_{i'=1}^{n} k'_i}, \tag{4.1}$$

onde n é o número de nós na rede antes da adição do nó i. Para cada nó j sorteamos um número aleatório e caso este número seja menor que pl criamos uma ligação entre os nós i e j. Esta distribuição tem média um, mas o menor grau possível é $k_i = 1$, então caso o novo nó não receba nenhuma ligação, este não é adicionado. Portanto o número médio de ligações dos nós adicionados por criação é maior que um. Em uma medida posterior à construção da rede encontramos $\langle m \rangle \approx 3.6$.

Os nós adicionados por duplicação seguem um modelo semelhante ao de Duplicação-Divergência (2.3.2) que representa a duplicação e subsequente subfuncionalização de genes. Entretanto, consideramos que assim como a topologia da rede influencia a escolha dos vizinhos no processo de criação, ela deve ter importante papel na duplicação dos nós. As redes de associação proteica possuem um grande número de módulos funcionais, que são estruturas formadas por um grande número de nós altamente ligados entre si e que participam de uma mesma função celular [23]. É razoável, então, assumir que uma alteração em um nó pertencente a um módulo funcional causará um grande impacto na função à qual pertence. A propriedade topológica que caracteriza esses módulos são um altos graus e clusterização. De fato, alguns trabalhos apontam correlações negativas entre duplicabilidade e grau, e duplicabilidade e clusterização [35–37].

O processo de duplicação será implementado da seguinte forma:

- um nó i é escolhido para ser duplicado segundo uma probabilidade de duplicação dada por

$$pd_i = \frac{(1 - C_i)}{k_i} \left[\sum_{j=1}^{n} \frac{(1 - C_j)}{k_j} \right] \; ; \tag{4.2}$$

- é adicionado um nó i' com todos os vizinhos do nó i, e com uma ligação entre eles, para evitar redes bipartidas;

- para cada nó j ligado à i e i', escolhemos aleatóriamente uma das duas ligações e eliminamos com probabilidade q;

O nosso modelo conta então com dois mecanismos de adição de nós, duplicação e criação. Definiremos então um parâmetro p que determinará a quantidade de nós adicionados por criação,

de forma que o número de nós provenientes desse método seja o produto de p pelo número total n de nós na rede $p \times n$. Logo o número de nós adicionados por duplicação é $n(1-p)$. Dada uma rede, o modelo sorteia o mecanismo de adição do próximo nó, dado pelo parâmetro p, e o adiciona, segundo os métodos apresentados acima, até termos n nós. Temos, então, dois parâmetros livres no modelo, o parâmetro p que determina a proporção de nós adicionados por cada método, e o parâmetro q, herdado do modelo de Duplicação-Divergência.

Capítulo 5

Resultados

Na figura 5.1 temos uma comparação das medidas para uma rede obtida através do modelo com parâmetros $p = 0.1$ e $q = 0.05$ e as redes dos organismos apresentadas na figura 3.2. Primeiramente podemos observar na figura **5.1 c** que a rede tem um grau médio aproximadamente igual ao dos organismos tomados como padrão. Encontramos uma distribuição de grau semelhante ao das redes de associação proteica, principalmente o comportamento para graus intermediários e o aumento da distribuição para alto grau, apesar de encontrarmos um decréscimo para graus pequenos. O modelo reproduz também uma rede altamente clusterizada e com um número médio de vizinhos por grau muito semelhante às redes dos organismos padrão. A figura 5.2 mostra uma comparação entre os resultados obtidos com os três modelos para reprodução das redes de associação protéica.

Comparando as matrizes de associação ordenadas para os organismos e para os modelo na figura 5.3, vemos que os modelo de Barabási-Albert e Duplicação-Divergência produzem redes com matrizes de associação mais estreitas, ou seja, que a distância do nó até seus vizinhos na matriz de associação é menor que os organismos padrão, e nenhum destes modelos resulta na formação de módulos. A rede de Barabási-Albert possui uma região nos extremos onde os vizinhos estão todos a uma certa distância, gerando um espaço vazio. O modelo Híbrido resulta em uma matriz de associação com uma forma de "folha" semelhante às das redes de associação proteica. Além disso, temos um módulo altamente ligado no centro do ordenamento.

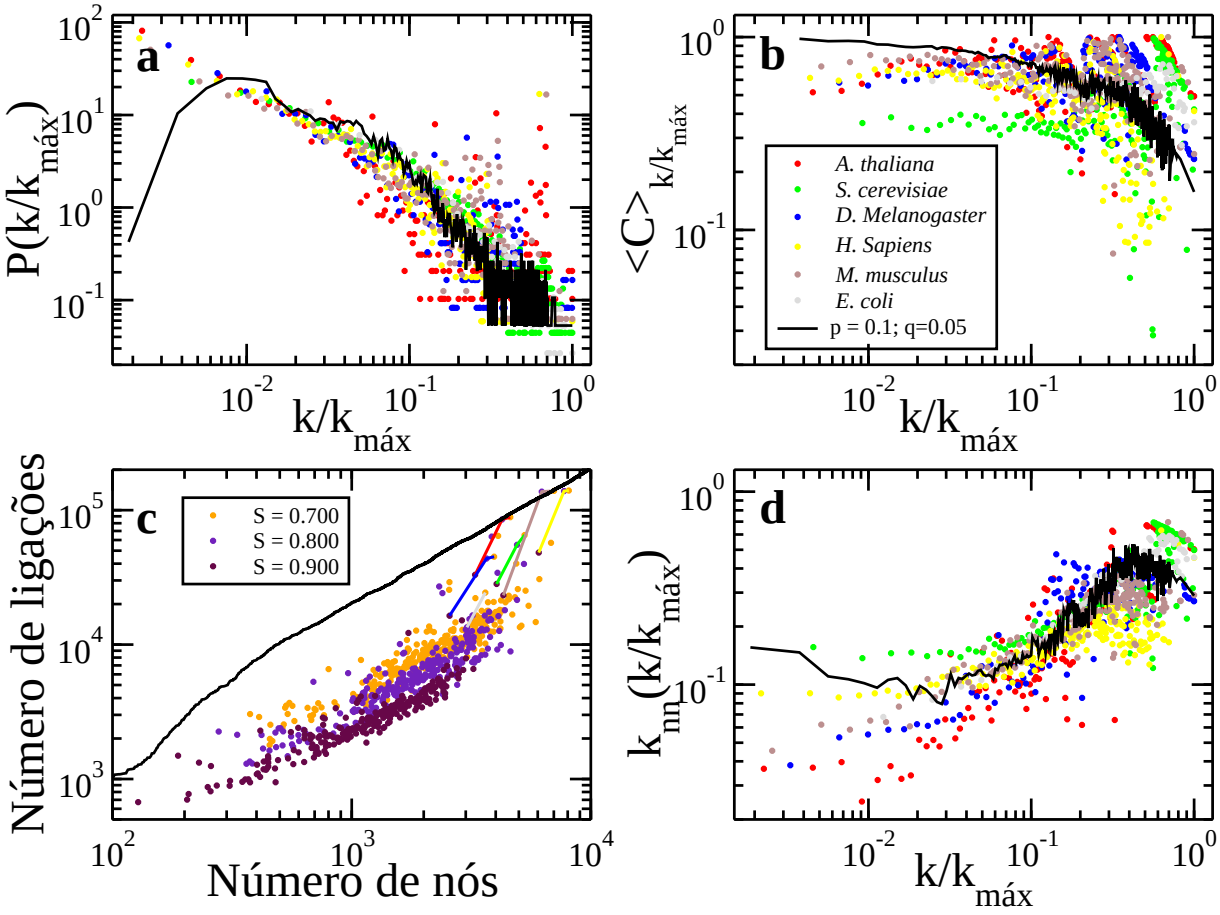

Figura 5.1: Comparação da rede obtida com o modelo e das redes dos organismos padrão *Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster, Homo sapiens, Mus musculus* e *Escherichia coli.*

A seguir iremos comparar as redes obtidas com diferente parâmetros do modelo, lembrando que no limite de $p = 1$ recuperamos o modelo de Barabási-Albert, e com $p \approx 0$ e q alto recuperamos o modelo de Duplicação-Divergência. Os pontos em cinza são as redes de todos os organismos *core* do STRING para índice de confiança $S = 0.800$. Na figura 5.4 fixamos o parâmetro $q = 0.05$ e variamos o parâmetro p. À medida que p aumenta a rede se aproxima de uma rede obtida pelo modelo de Barabási-Albert. Apesar de explorarmos apenas até $p = 0.5$, onde metade dos nós são adicionados por cada método, vemos que a distribuição de grau já não apresenta uma elevação para alto grau, bem como a clusterização diminui, e a rede perde o seu caráter assortativo. Esses resultados são compatíveis com o modelo de Barabási-Albert.

Na figura 5.5 mantivemos $p = 0.1$ e variamos o parâmetro q. As redes começam a se aproximar das redes obtidas através do modelo de Duplicação-Divergência, a distribuição de grau se torna

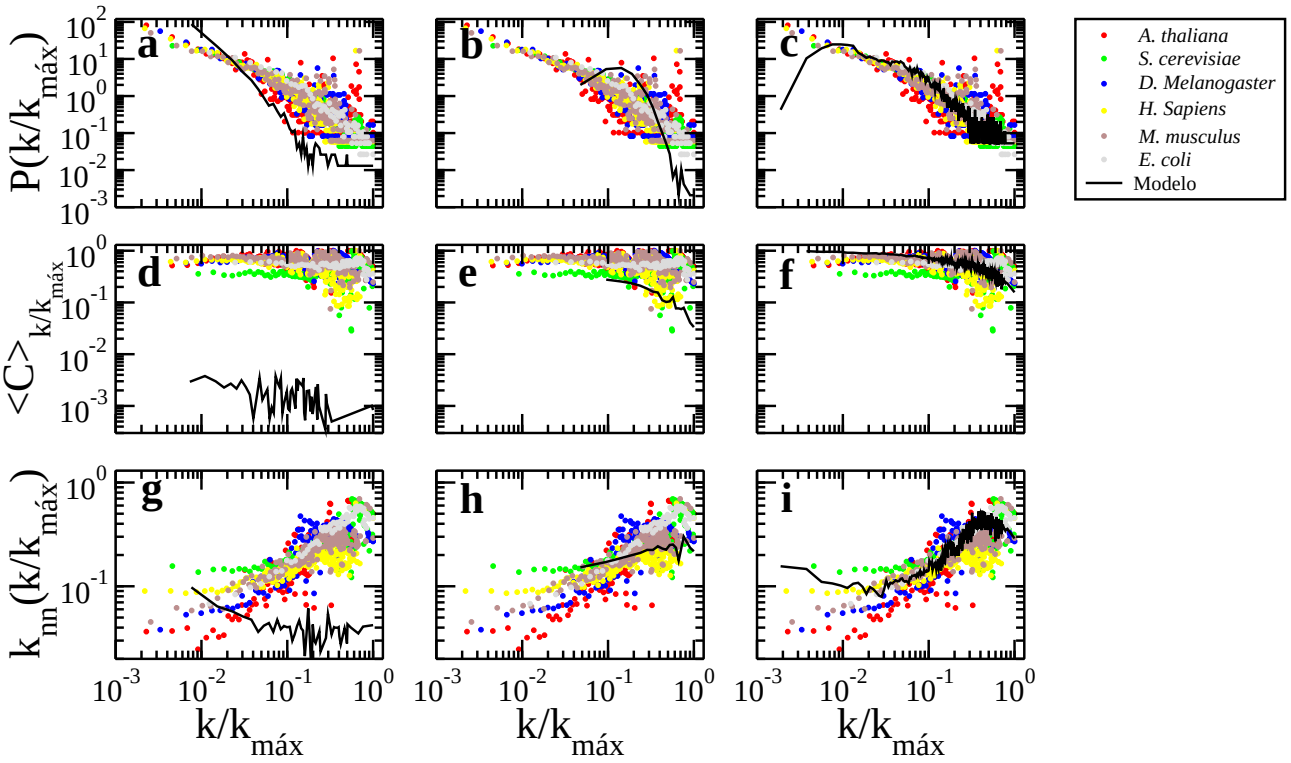

Figura 5.2: Comparação entre os resultados obtidos com os três modelos. Na primeira coluna temos o modelo de Barabási-Albert, Duplicação-Divergência na segunda, e os resultados do nosso modelo na última, com os parâmetro $p = 0.1$ e $q = 0.05$.

uma lei de potência, a clusterização diminui e a rede perde o caráter fortemente assortativo.

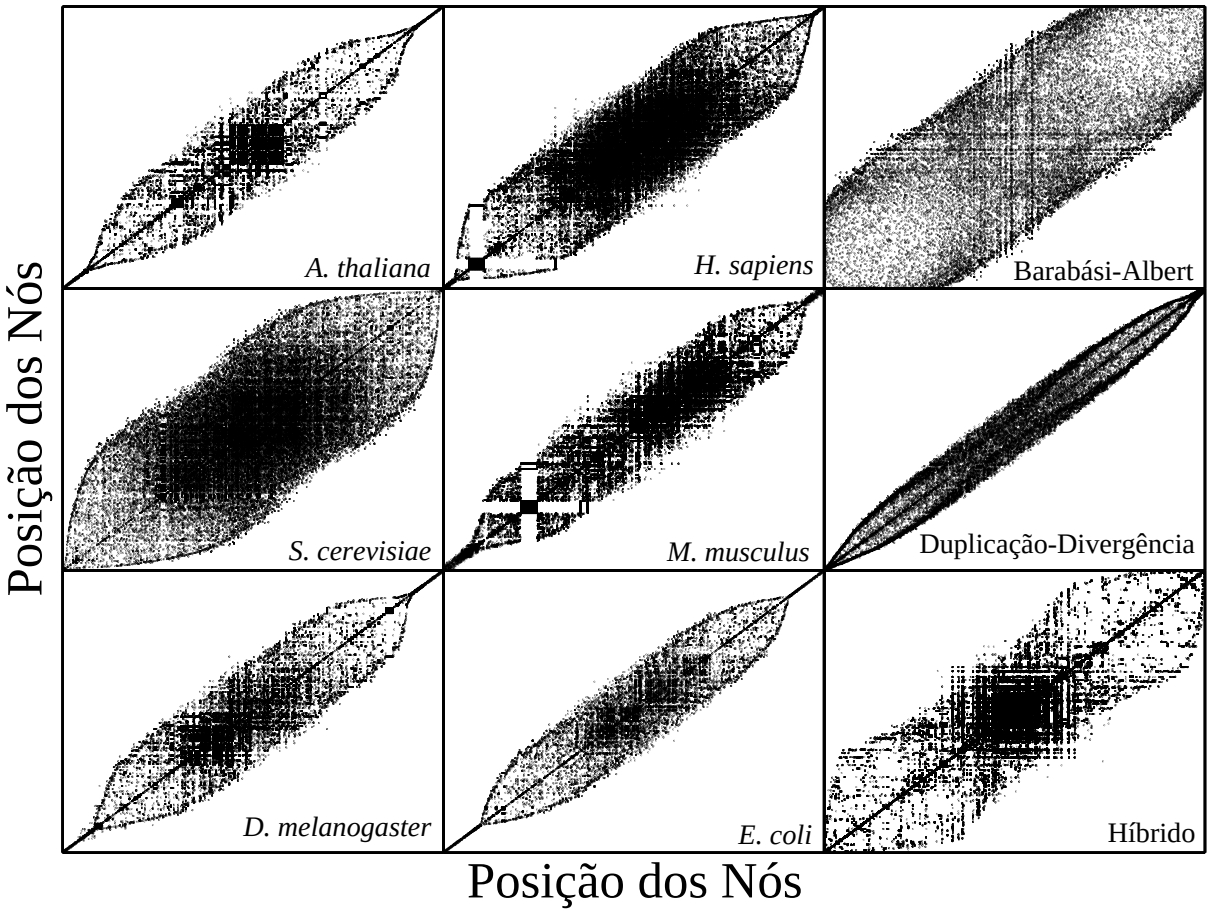

Figura 5.3: Matrizes de associação ordenadas para os organismos padrão e modelos.

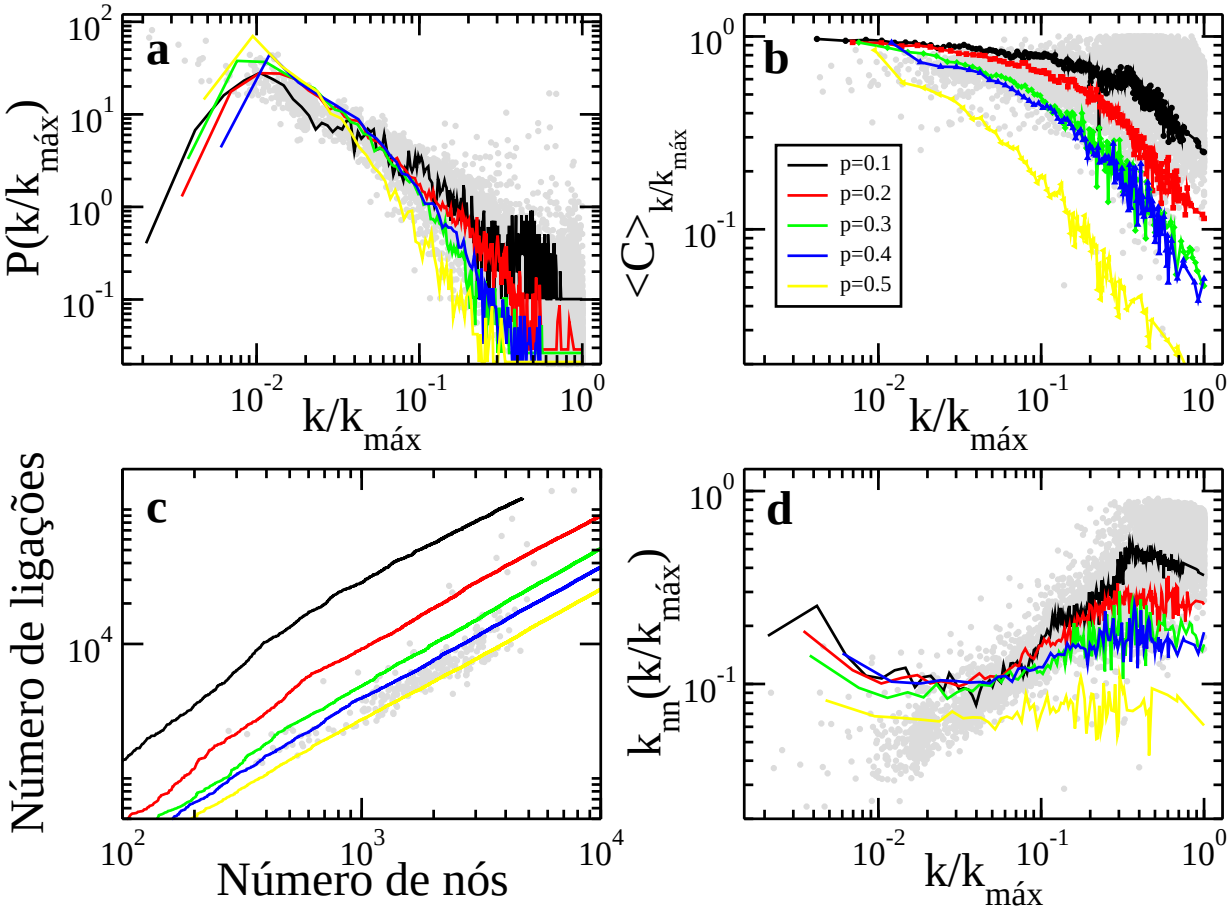

Figura 5.4: Redes obtidas variando o parâmetro p e mantendo $q = 0.05$. Para comparação mantemos as redes dos organismos para S=0.800 em cinza.

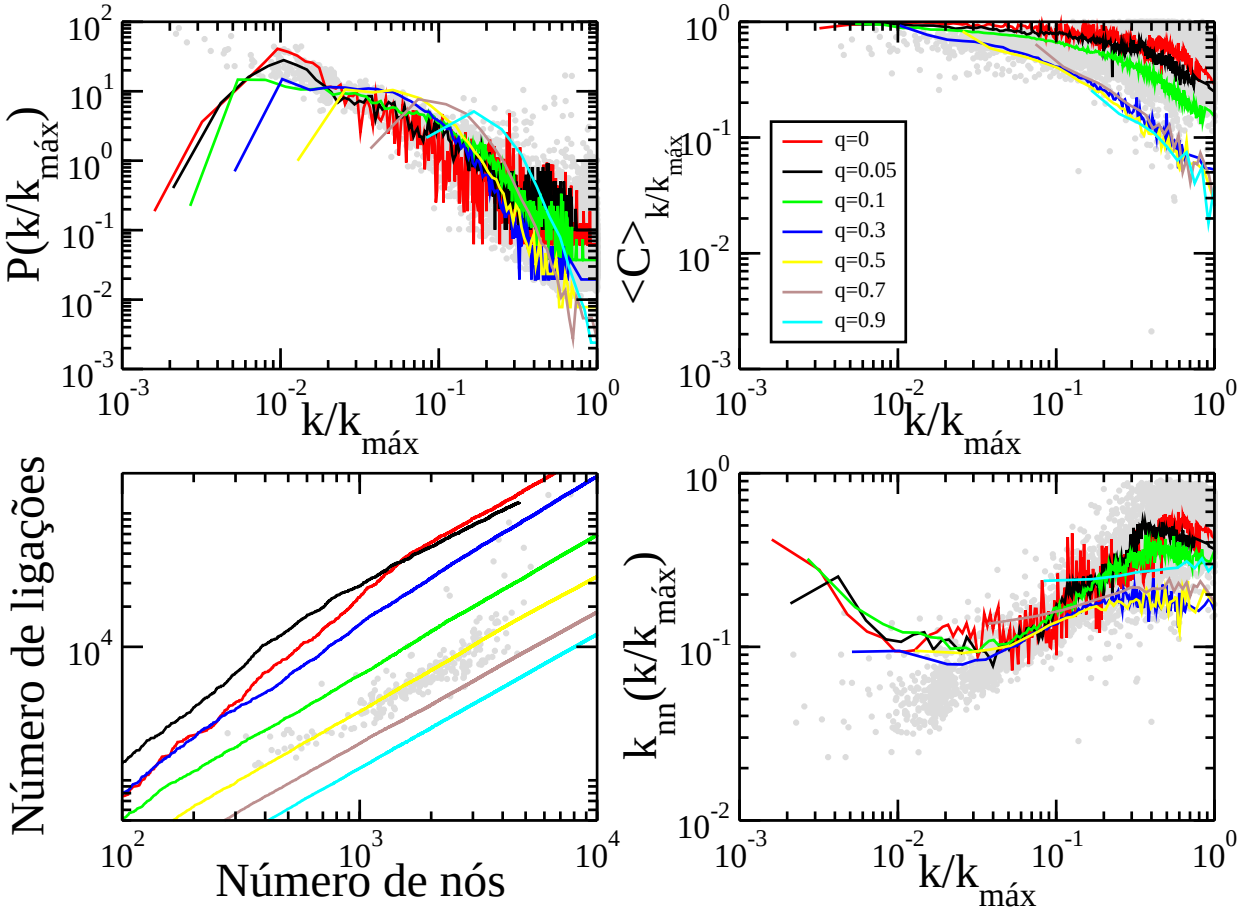

Figura 5.5: Redes obtidas variando o parâmetro q e mantendo $p = 0.1$.Para comparação mantemos as redes dos organismos para S=0.800 em cinza.

Capítulo 6

Discussão e Conclusões

Neste trabalho conseguimos propor um modelo de crescimento de redes que reproduz, com razoável fidelidade, as redes de associação proteica, cuja regra de adição de nós foi baseada nos mecanismos biológicos de crescimento do genoma.

Em primeiro lugar podemos destacar que encontramos uma distribuição de grau com um decaimento na forma de lei de potência para graus intermediários, e principalmente, com uma elevação para grau mais alto. Apesar do nosso modelo ser composto de generalizações de dois modelos anteriores, e algumas propriedades estarem presente em um ou outro modelo, este comportamento para grau mais alto não pode ser explicado independentemente por nenhum dos dois. Ainda que tenhamos encontrado uma distribuição de grau abaixo do esperado para os nós com grau pequeno, temos que considerar que estas proteínas tem apenas algumas ligações. Isto significa que estas, em particular, podem não ter sido tão profundamente estudadas, o que pode acarretar em um grande número de falsos negativos. Assim, caso algumas dessas proteínas tenham um número maior de ligações, a distribuição de grau diminui para grau pequeno. Outra propriedade interessante do modelo é a geração de redes altamente clusterizadas, mesmo escolhendo nós pouco clusterizados para sofrer duplicação. Isto ocorre pois, como o parâmetro q, que representa a mutação, é baixo, o par duplicado forma um triângulo para cada vizinho que não perde nenhuma ligação. Encontramos ainda uma rede assortativa, provavelmente herança do modelo de Duplicação-Divergência. O resultado do algoritmo de ordenamento indica que estamos

produzindo uma rede que reproduz estruturas de vizinhança presentes nas redes de associação proteica.

Em particular, podemos destacar que a proporção de genes adicionados por duplicação, isto é, cerca de 90%, é condizente com valores encontrados na literatura [2]. Isto é mais um indício da validade biológica das hipóteses utilizadas para construir o modelo. Outra característica interessante do modelo é a proposta de uma correlação entre a topologia local da rede em determinado nó, e a probailidade de este sofrer duplicação, no caso atual a probabilidade de duplicação é proporcional à $(1 - C_i)/k_i$. Isto permite que testemos diferentes relações topológicas, ganhando assim uma intuição do fenômeno biológico de duplicação de genes.

Em futuros trabalhos, poderemos explorar mais a fundo o impacto da topologia na duplicação de nós, por exemplo adicionando expoentes, de forma que possamos controlar o impacto da clusterização e do grau separadamente, ou ainda propor diferentes topologias. Nesta linha podemos também analisar as redes de diversos organismos buscando indícios da topologia subjacente ao processo de duplicação de genes. Poderemos também buscar outras ferramentas utilizadas para caracterizar redes e aplicá-las às redes de associação protéica e ao nosso modelo, por exemplo um algorítmo de ordenamento em duas dimensões, bem como caracterizar melhor os módulos gerados pelo ordenamento.

Apêndice A

Variação de parâmetros dos modelos de Barabási-Albert e Duplicação divergência

Na figura A.1 temos as redes obtidas através do modelo de Barabási-Albert para valores diferentes de m com $m_0 = 5$. Podemos ver que a escolha de parâmetros não produz grandes variações na rede.

Na figura A.2 temos as redes obtidas variando o parâmetro q no modelo de Duplicação-Divergência. Vemos que todos os parâmetros resultam em redes com baixa clusterização e assortativas, mas que para um parâmetro q baixo a rede perde completamente o comprtamenteo de lei de potência e a distribuição de grau adquire uma foram de arco.

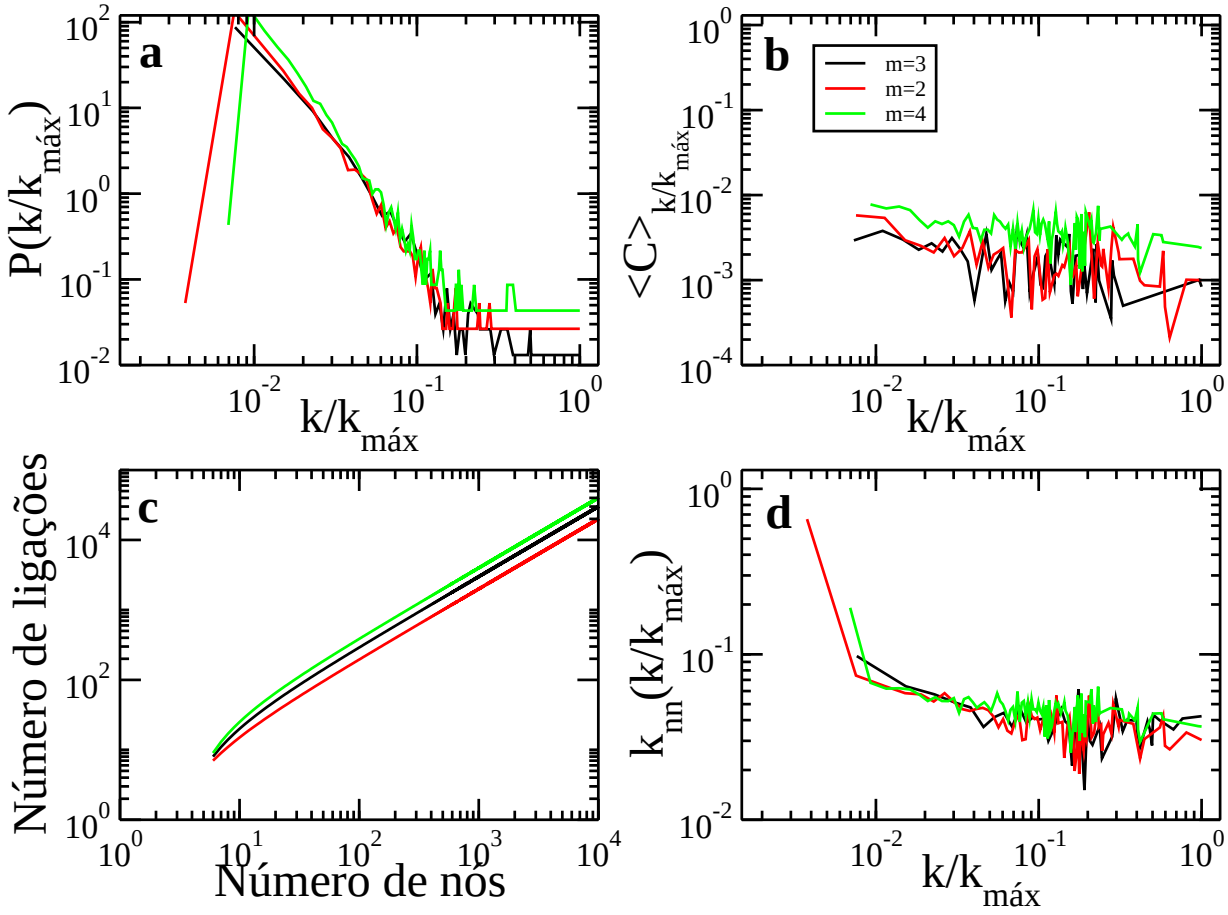

Figura A.1: Redes obtidas do modelo de Barabási-Albert variando o parâmetro m.

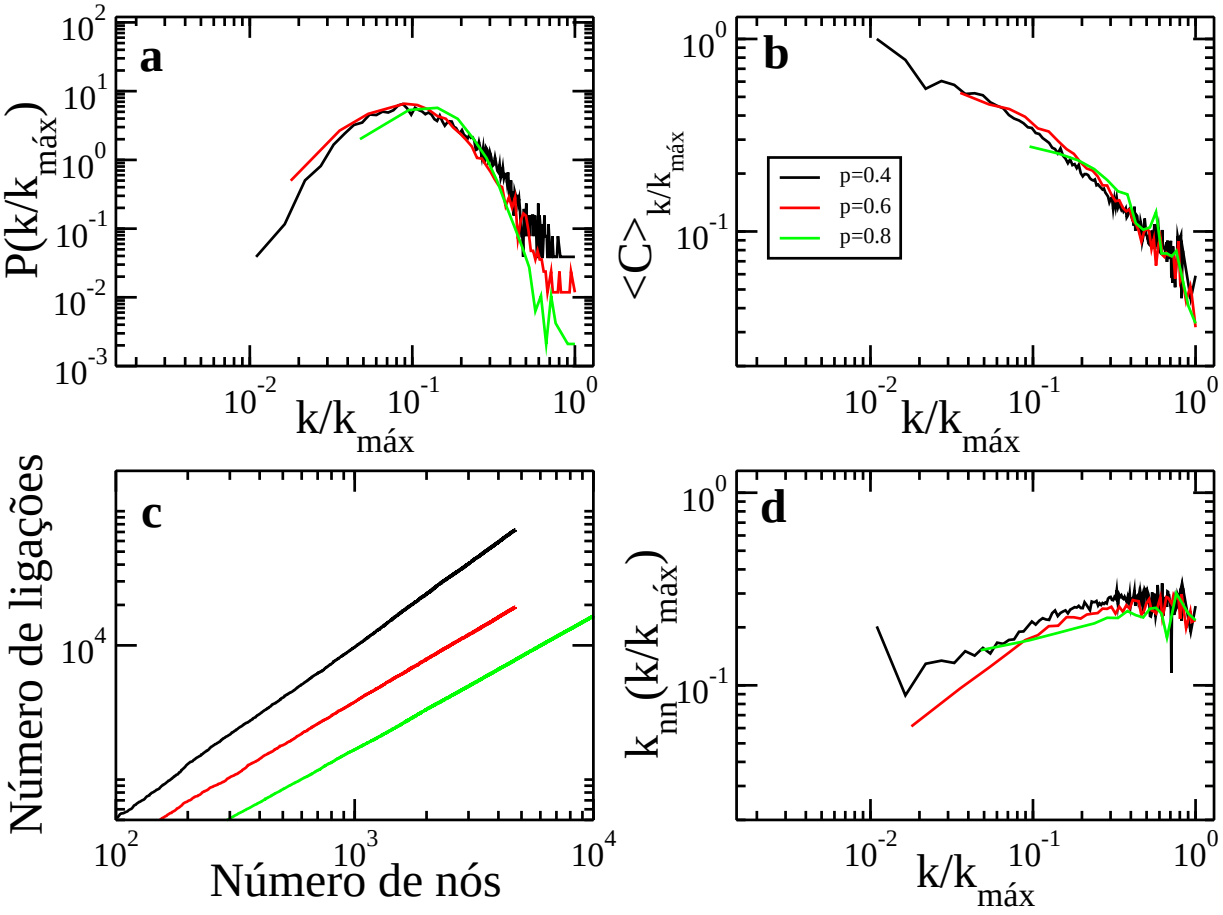

Figura A.2: Redes obtidas do modelo de Duplicação-Divergência variando o parâmetro q.

Referências Bibliográficas

[1] Barabási, A.-L. and Albert, R., Science **286** (1999) 509.

[2] Zhou, Q., Zhang, G.-j., Zhang, Y., and Spring, C., Genome Research (2008).

[3] Darwin, C., *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.*, London, 1859.

[4] Mendel, J. G., Verhandlungen des naturforschenden Vereines in Brünn (1866).

[5] Brown, T. A., *Genomes*, John Wiley & Sons, 2 edition, 2002.

[6] Castro, M. A. A., Dalmolin, R. J. S., Moreira, J. C. F., Mombach, J. C. M., and de Almeida, R. M. C., Nucleic acids research **36** (2008) 6269.

[7] Long, M., Betrán, E., Thornton, K., and Wang, W., Nature reviews. Genetics **4** (2003) 865.

[8] Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J., Proceedings of the National Academy of Sciences **103** (2006) 9935.

[9] Heinen, T. J. A. J., Staubach, F., Häming, D., and Tautz, D., Current biology : CB **19** (2009) 1527.

[10] Brosius, J., Science **251** (1991) 753.

[11] Force, A. et al., Genetics **151** (1999) 1531.

[12] Innan, H. and Kondrashov, F. A., Nature reviews. Genetics **11** (2010) 97.

[13] Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V., Genome biology **3** (2002).

[14] Lynch, M. and Force, A., Genetics **154** (2000) 459.

[15] Lynch, M., Science **297** (2002) 945.

[16] Koonin, E. V., Makarova, K. S., and Aravind, L., Annu. Rev. Microbiol **55** (2001) 709.

[17] Ochman, H., Lawrence, J. G., and Groisman, E. A., Nature **405** (2000) 299.

[18] Keeling, P. J. and Palmer, J. D., Nature reviews. Genetics **9** (2008) 605.

[19] Newman, M. E. J., SIAM Review **45** (2003) 167.

[20] Albert, R. and Barabási, A. L., Reviews of modern physics **74** (2002) 47.

[21] Colizza, V., Flammini, A., Maritan, A., and Vespignani, A., Physica A: Statistical Mechanics and its Applications **352** (2005) 1.

[22] Costa, L. F., Rodrigues, F. A., Travieso, G., and Boas, P. R. V., Advances in Physics **56** (2007) 167.

[23] Rybarczyk-Filho, J. L. et al., Nucleic acids research (2010) 1.

[24] Kirkpatrick, S.y Gelatt, C. J. and Vecchi, M., Science **220** (1983) 671.

[25] Bansal, S., Khandelwal, S., and Meyers, L. A., BMC bioinformatics **10** (2009) 405.

[26] Evlampiev, K. and Isambert, H., BMC systems biology **1** (2007) 49.

[27] Mithani, A., Preston, G. M., and Hein, J., Bioinformatics (Oxford, England) **25** (2009) 1528.

[28] Takemoto, K. and Oosawa, C., Mathematical biosciences **208** (2007) 454.

[29] Barabási, A.-L. and Oltvai, Z. N., Nature reviews. Genetics **5** (2004) 101.

[30] Vázquez, A., Physical Review E **67** (2003) 1.

[31] Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A., Complexus **1** (2003) 38.

[32] von Mering, C. et al., Nucleic acids research **33** (2005) D433.

[33] von Mering, C. et al., Nucleic acids research **35** (2007) D358.

[34] Jensen, L. J. et al., Nucleic acids research **37** (2009) D412.

[35] Prachumwat, A. and Li, W.-h., Molecular Biology **7**.

[36] Li, L., Huang, Y., Xia, X., and Sun, Z., Molecular biology and evolution **23** (2006) 2467.

[37] Dalmolin, R. J. S. et al., Preprint .